IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

A modeling framework for generation of positional and temporal simulations of transcriptional regulation

David A. Knox, *Member, IEEE,* and Robin D. Dowell, *Member, IEEE*

Abstract—We present a modeling framework aimed at capturing both the positional and temporal behavior of transcriptional regulatory proteins in eukaryotic cells. There is growing evidence that transcriptional regulation is the complex behavior that emerges not solely from the individual components, but rather from their collective behavior, including competition and cooperation. Our framework describes individual regulatory components using generic action oriented descriptions of their biochemical interactions with a DNA sequence. All the possible actions are based on the current state of factors bound to the DNA. We developed a rule builder to automatically generate the complete set of biochemical interaction rules for any given DNA sequence. Off-the-shelf stochastic simulation engines can model the behavior of a system of rules and the resulting changes in the configuration of bound factors can be visualized. We compared our model to experimental data at well-studied loci in yeast, confirming that our model captures both the positional and temporal behavior of transcriptional regulation.

Index Terms—Biological system modeling, automated model building, transcription regulation modeling.

1 Introduction

Transcriptional regulation is the system behavior arising from the interaction of numerous regulators with DNA. This complex system produces precise gene expression at specific times and locations. Experimental studies of gene expression have unlocked the function of many proteins involved in regulating the transcription process [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. New experimental techniques are being developed to understand transcriptional regulation at unprecedented temporal and molecular detail, ultimately even at single-cell resolution [12], [13], [14], [15]. Yet much is still to be learned.

There is growing evidence that transcription emerges not solely from the individual components, but rather from the collective behavior between the components [15], [16], [17], [18], [19], [20], [21]. Three major classes of protein regulators: transcription factors, nucleosomes, and the transcriptional machinery, interact with DNA in both competitive and cooperative fashion. DNA undergoes millions of interactions every second, constantly changing the configuration of the molecular components bound. It is the stochastic, temporal, and spatial interactions of these regulators that control the transcription process in each individual cell.

Encapsulating our understanding of these interactions into a computational model is integral to understanding transcriptional regulation [22]. Models allow us to explore a system, create testable hypotheses, and identify when key details are missing in our current knowledge. To date, most modeling frameworks have either focused on the detailed molecular behavior of a specific regulator or the interaction of a small subset of regulatory components [8], [23], [24], [25], [26], [27], [28]. Yet, few models have approached the problem of simultaneously capturing the behavior of all three major regulator classes. In part, this is because most models either focus on the positional information of each component [8], [21] or the temporal behavior of their inherent dynamics [26], [29]. Integrating both the positional information and temporal information often leads to computationally expensive models.

Yet, as experimental techniques continue to improve, modeling approaches must also evolve to represent increasingly realistic molecular details while still remaining computationally tractable. We seek to construct biologically realistic computational models that capture not only the positional binding of transcription factors and nucleosomes, but also the underlying temporal dynamics, such as the behavior of transcriptional machinery during initiation and elongation. Therefore, we have developed a modeling framework that can automatically generate rule sets describing the possible molecular interactions implied by a given DNA molecule. In this work, we describe the basic biology of regulation (section 2), the design of our modeling framework (section 3), case studies that demonstrate the validity of our model (section 4), and discuss its potential and limitations (section 5).

2 Biological Background

This section reviews the basic concepts and components of transcriptional regulation that are necessary to understand our system. Readers familiar with the biology may safely skip this section. For more information on the transcription process, please refer to [30], [31], [32].

D.A. Knox is with the Computational Bioscience Program at the University of Colorado, School of Medicine, Anschutz Medical Campus, Aurora, CO 80029. E-mail: david.knox@colorado.edu.

R.D. Dowell is with the Molecular, Cellular, and Developmental Biology Department, BioFrontiers Institute, University of Colorado, Boulder, CO 80309. E-mail: robin.dowell@colorado.edu.

Manuscript received February, 2014; revised July 8, 2015.

Digital Object Identifier no. 10.1109/TCBB.2015.2459708

2.1 Overview of Transcriptional Regulation

Transcription is the process of copying stretches of DNA into RNA, the necessary first step of all cellular processes. Therefore, transcription is a critical aspect of all cellular activities. Understanding how transcription is regulated, namely when (temporal) and where (positional) RNA is produced, is the underlying goal of transcriptional modeling systems.

2.2 Components of regulation

DNA is the central molecule of transcriptional regulation. Whereas the concentration of all other components varies based on condition or cell type, the number of copies of the DNA per cell is largely defined by the organism. The DNA encodes the instructions for when, where, and how much of each transcript is produced.

Three major classes of DNA binding proteins are involved in transcriptional regulation: transcription factors [6], [33], [34], nucleosomes [35], [36], [37], and the transcriptional machinery [6], [38]. Most DNA binding proteins recognize specific sequences of DNA with different affinities. A position specific scoring matrix (PSSM) is typically utilized to describe the sequence affinity preferences of each DNA binding protein [39]. The physical binding of a regulator to DNA depends on not only its PSSM, but also its cellular concentration. At higher concentrations, the best matches to the PSSM will become saturated and the protein is more likely bind to lower affinity sites [40].

In addition, a regulator's binding to DNA may be influenced by competition with other proteins for a sequence, the nearby positioning of nucleosomes, the presence of cofactors, and the post-translational state of the protein itself. Particular configurations of interacting molecules are necessary for the recruitment of the transcriptional machinery and activation of transcription. Therefore, these proteins interact in complex ways, both cooperatively and competitively, to bind to DNA and induce transcription.

2.3 Dynamics of regulation

Recent experimental work has highlighted a number of inherently dynamic events that contribute to transcriptional regulation. The interaction of a regulatory protein with DNA is transient, as these factors are thought to bind and release frequently [41]. Transcription factor residency times (how long a factor binds at an individual site) may be an important but previously overlooked aspect of regulation [11]. The histone proteins within a nucleosome are modified, swapped, or displaced during transcription, which changes the behavior of individual nucleosomes [42], [43]. Finally, the movement of the transcriptional machinery along DNA is a highly dynamic process that pauses, shows variable processivity, and likely evicts DNA binding factors that impede its forward progress [44]. When one transcriptional machine directly impacts a second transcriptional process, this interaction is referred to as transcriptional interference [45]. Transcriptional interference has been shown experimentally to be a critically important aspect for the regulation of some loci [46], [47], [48], [49] .

The dynamic aspects of regulation are most apparent when examining single cell transcription. Recent technological innovations, such as fluorescent protein tracking and real-time nascent transcription observations [14], [50], have made single cell molecular behaviors more apparent. In these studies, variability in transcription and protein expression is widely observed, likely stemming from fluctuations in cellular abundances of proteins, the stochastic nature of molecular interactions, and microenvironments within a cell [51], [52]. Transcriptional regulation is inherently dependent upon the biochemical interactions of many different molecules, but robust enough to handle the stochastic fluctuations inherent in a molecular system. The resulting cell-to-cell variability is likely fundamental to most, if not all, molecular cellular processes [53], [54] .

3 Modeling Framework

A wide variety of techniques have been previously applied to the problem of modeling transcriptional regulation [55]. It is instructive to briefly review the existing modeling approaches (section 3.1) before discussing the overall goal and approach of our model (section 3.2). We then describe our representational framework (section 3.3) and provide details of our implementation (section 3.4).

3.1 A brief comparison of modeling strategies

Models of transcriptional regulation vary tremendously in their underlying level of abstraction of the biological process. At one end of the modeling spectrum are gene regulatory networks that seek to capture the logic of a circuit by describing the behavior between genes. At the other end, molecular dynamics focuses on the physics behind the atomic interactions between molecules. Several recent reviews discuss the tradeoffs inherent in choosing an abstraction [55], [56], [57]. These modeling methods have been reviewed in more detail elsewhere, see [58] for review on inferring gene regulatory networks, [57] for general mathematical modeling methods, and [59] for molecular dynamics models. Most current techniques fall between these two extremes and focus either on the positional details [8], [21], [28] or the temporal dynamics [60], [61], [62] of the system of interest.

Configuration based modeling systems have been developed (concepts reviewed in [16]) to describe the positional binding configuration of proteins using either probabilities or thermodynamics. Currently, these models consider only nucleosome positioning and transcription factor binding. With these two components, a given nucleotide can be unbound, bound in a nucleosome, or bound in a particular transcription factor (Fig. S1). Conceptually, these states define a model where the transitions between the states depend on both the affinity of the component to the DNA sequence and its concentration [21]. As both nucleosomes and transcription factors bind to multiple nucleotide positions, the actual connectivity between states varies depending on the identity of the component in order to capture the specificity and length of binding. This is captured elegantly by a hidden Markov model (HMM) and allows large DNA sequences to be quickly modeled on conventional computer

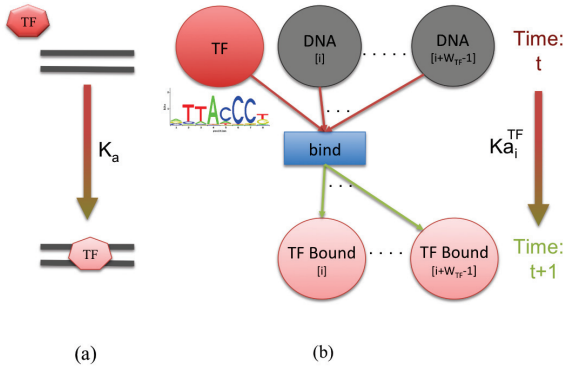

Fig. 1. Biological behaviors are modeled by computational components. (a) The core of our modeling framework is to capture biological interactions, such as the behavior of a transcription factor (TF) binding to the DNA (at rate K_a). (b) Abstract computational description of the biological interaction using Petri net notation. These descriptions are action based (blue rectangle) with the preconditions (circles listed above the action for a molecule of the TF (red) and multiple positions ($i...i + W_{TF} - 1$) of unbound DNA (grey)) required to apply the action, and the post-conditions (listed below the action with DNA positions ($i...i + W_{TF} - 1$) bound by that TF (pink)) that are true after the action is applied at time t. The position specific scoring matrix (PSSM, example shown below the TF as a sequence logo [67]) describes a specific TF's affinity for DNA, which is used to calculate the TF specific rate depending on the actual DNA being bound (Ka_i^{TF}). All action description are provided in Appendix 1.

resources [8], [21]. These models have proven to be quite successful at elucidating key regulatory principles inherent in the competition between nucleosome and transcription factors [16]. However, these methods describe only the population averaged behavior of a DNA region and do not address the inherent temporal variation in configurations within a single cell.

Temporal dynamics models seek to capture the key molecular interactions occurring during transcriptional regulation. Typically, these models describe the behavior of molecules through a series of biochemical rules (reviewed in [55]). Each rule specifies the reactants being combined at a specified rate to produce the resultants (Eq. 1). In many cases, the Gillespie stochastic simulation algorithm (SSA) is then applied to explicitly capture every interaction in a discrete and stochastic simulation [63]. The algorithm is a dynamic Monte Carlo method that stochastically simulates a set of biochemical reactions to produce one possible trajectory of the system. This approach has been used to model a variety of stochastic systems, from ecosystems to cells [64], [65]. Furthermore, these models have been integral to our understanding of how the act of transcription in one region can regulate nearby and overlapping transcription through interference [66].

3.2 Stochastic Models of Transcription

We sought to construct a modeling framework that integrates the sequence dependent positional information of DNA binding proteins with the inherent temporal dynamics of transcriptional interference. The configuration based modeling paradigm can capture steady state positional interactions, but must be extended to capture dynamic events, such as the movement of the transcriptional machinery. These dynamic events are not only positionally dependent, but also temporally dependent. For example, consider polymerase traversing DNA in the process of transcription. The location of polymerase is time dependent and its ability to move along the DNA is influenced by the state of the nucleotides ahead [68]. Capturing the temporal and positional dimensions simultaneously within the HMM framework leads to an explosion of alternative states, because each possible movement in the temporal dimension impacts multiple positional states (Fig. S2). Consequently, we sought an alternative representation that simplifies capturing both dimensions.

A new class of stochastic models of gene regulation are just emerging [26], [29], [69], [70] that are specifically tailored to deal with complex configurations of regulators at individual loci within single cells. Previous work has focused on detailed models of the transcriptional machinery in bacteria [26], [29], [69], [70]. For example, a discrete, stochastic model of biological elongation by the transcriptional machinery was developed to explore stochastic phenomena, such as the relationship between regulatory protein site occupancy and production rate [71]. Their models were described in a Petri net formalism that highlights both the discrete nature of molecular interactions and the dependencies within interaction networks. Similar models have even been extended to include translational output in bacteria where transcription and translation processes are intimately coupled [69].

These models capture the dynamics of individual molecular components, such as the transcriptional machinery, by taking advantage of the fact that all biological interactions can be described as chemical reactions [26], [69], [70]. Interactions between the DNA and binding factors are specified as a set of reactants that produce a set of resultants at a given rate (Eq. 1). These models seek to capture the molecular interactions occurring along DNA (positional) throughout time (temporal) using SSAs.

$$reactant + ... \xrightarrow{Ka} resultant + ... \quad (1)$$

We sought to develop a similar framework for eukaryotic transcription, where nucleosomes introduce an added layer of complexity to transcriptional regulation. Therefore, our model integrates the basic behaviors of transcriptional machinery, similar to earlier stochastic models, with the sequence specificity of nucleosome and transcription factor binding, as is more typically captured by hidden Markov formulations. This synthesis can be captured by the chemical reaction based stochastic simulation approach. Additionally, we sought to automate the model generation process so that a sequence specific collection of reaction rules could be created quickly and systematically. Our framework describes interactions using abstract rules (described as Petri nets) and the creation of a specific model by the application of those abstract rules to a given DNA sequence. However, there is a big drawback to typical stochastic simulation models. They are computationally expensive, as every molecule requires rules to describe the set of all possible interactions. This leads to a combinatorial explosion of possibilities as the number of components in the model increases. This is particularly true for the central molecule of transcriptional regulation: DNA.

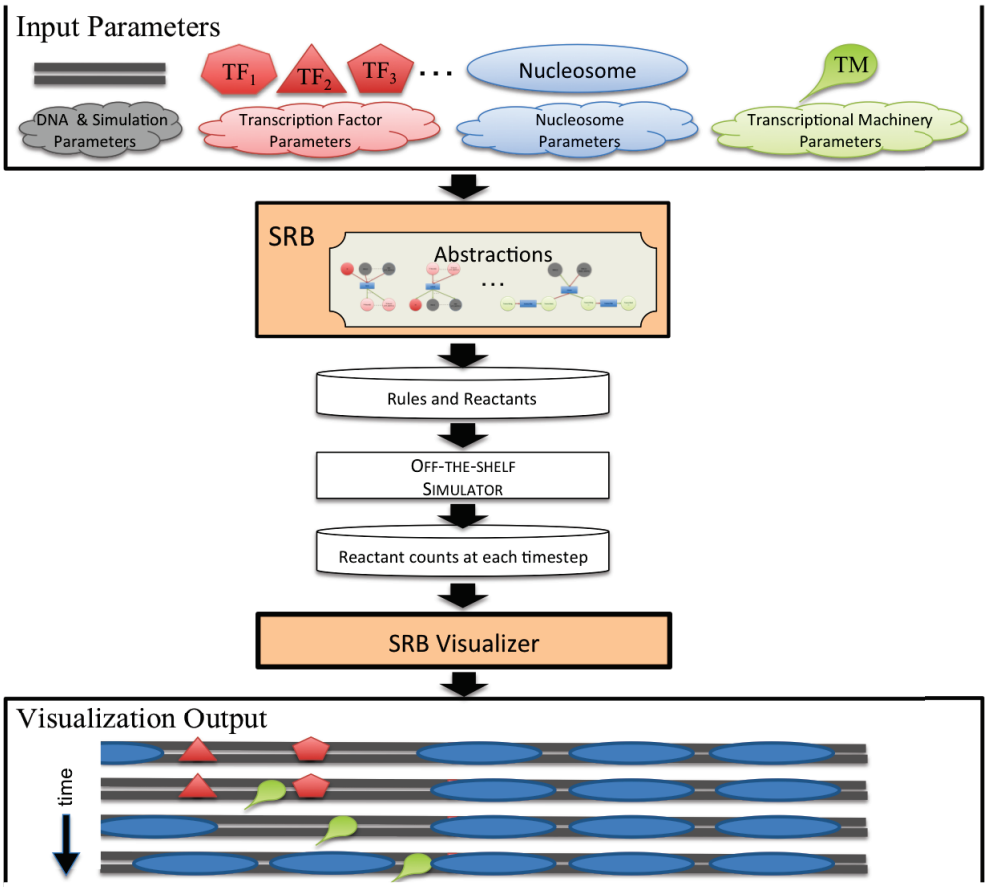

Fig. 2. Flowchart depicting the Stochastic Rule Builder (SRB) and visualization pipeline. The SRB encapsulates the abstractions (Petri net descriptions, Appendix 1) for all the interactions and applies those interactions to specific sequences of DNA (grey bars) using the user provided component parameters (input parameters). Output of the SRB is both a set of reactants representing all the different states of the components and the rules using those reactants. The rules are simulated using an off-the-shelf simulation engine, which produces an output file containing each reactant's molecular count at each time step. We note that each time step spans many interactions and only the final system state is recorded in the results. The molecular counts are interpreted by our Visualizer to generate the configuration of the components along the DNA at each time step (visualizer output). Transcription factor (TF); Transcriptional Machinery (TM).

3.3 Representational framework

We first seek to address the problem of a combinatorial explosion of configurations. In reality, DNA is a chain of nucleotides. Inspired by the positional models of transcription, the key observation is that individual DNA binding proteins interact with a short contiguous string of bases. Because of steric constraints, only a single binding protein can interact with a given nucleotide at a given time. Therefore, by treating DNA not as a single molecule but as a string of entities, we can decouple the actions of one DNA binding protein from another, so long as they interact sufficiently far away from each other.

Our framework uses a rule-based methodology that defines all the interactions between the DNA and factors as chemical equations (Eq. 1). To this end, we use Petri nets to focus on describing processes from an event centric perspective (Fig. 1). This shifts the focus squarely to the dynamic temporal events permissible for a particular component, while still allowing the events to be decomposed into their local positional effect (Fig. S3). Stochastic Petri nets (SPN) have been utilized successfully to model diverse biological processes, including metabolic networks, signaling pathways, and gene regulatory networks [71], [72], [73], [74], [75]. Our use of the graphical representation creates easy to understand events and allows our framework to be highly flexible and extensible.

These abstract rules (described in Appendix 1) are then applied to a specific DNA sequence to obtain a complete set of model rules that can be simulated by off-the-shelf stochastic simulation engines. Each temporal event, such as binding, depends on the current state or local configuration of the DNA (preconditions). The positions of the DNA influenced by a rule (its span) are also specified. Execution of an event results in a change to the local configuration. We are able to describe the specific positions of an interaction, creating an abstract description of each event that is independent of other events. To create a runnable simulation, these abstract descriptions are applied to a specific DNA sequence to generate a set of rules defining the permissible molecular interactions. An additional advantage of using SSA simulations is it allows us to visualize the configuration of each position at each time step, as well as infer the dynamic movement of factors along the DNA (Fig. S4).

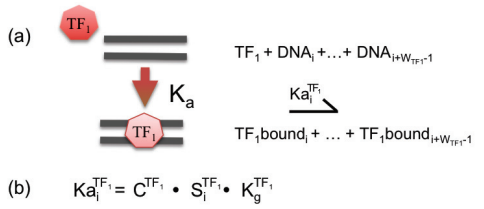

Fig. 3. Generation of the biochemical interaction rules requires interaction rates to be a function of the specific sequence. (a) The biochemical interaction rules generated by the binding of a specific transcription factor (TF_1) at position i of a specific DNA sequence (DNA_i + .. + $DNA_{i+W_{TF}-1}$) to produce bound factor (TF_1bound_i + .. + $TF_1bound_{i+W_{TF}-1}$). (b) The position and TF specific reaction rate ($\text{Ka}_i^{TF_1}$) is a function of the transcription factor concentration (C^{TF_1}), the transcription factor's affinity for a specific sequence of DNA ($S_i^{TF_1}$) and the generic association rate ($K_g^{TF_1}$) of the transcription factor.

3.4 Implementation

Our representational framework defines a set of interaction rules that are independent of the DNA being modeled. In this section, we describe how we apply the rules to a specific DNA sequence, simulate the resulting system of interactions, and analyze the results. The complete overview of this process is shown in Fig. 2.

3.4.1 Stochastic Rule Builder

The stochastic rule builder (SRB) converts our abstractions into a runnable set of rules specific to a particular DNA sequence. Given our framework, the creation of a specific executable model becomes a conversion process, similar to a compiler. The SRB takes as input the specific DNA sequence, a configuration file, and the necessary parameters to build a complete set of rules for the simulator (Fig. S5). SRB utilizes our generic action descriptions (the Petri nets; Appendix 1 and Tables S1-S3) to generate rules in the form of chemical reactions. Each rule specifies a set of reactants that are transformed at some rate (K_a) to produce a set of resultants (Fig. 1).

For each generic component, the SRB creates sequence specific rules. Consider a transcription factor as an example. A transcription factor has particular sequences to which it preferentially binds, described by a PSSM. This matrix specifies the number of positions, as well as the nucleotides that are preferentially bound at each position. At every position of a specified DNA sequence, the SRB utilizes the PSSM to calculate how well the transcription factor could bind. Fig. 3A shows an abstract rule for the binding of a transcription factor to the DNA. When a rule is generated for the binding of this transcription factor to a given DNA position, a binding rate (K_a) must be specified. A sequence specific function for the rate (Ka_i^{TF1}) of binding takes into account the factor's strength of binding the specific sequence ($S_i^{TF_1}$, inferred from the PSSM) and a generic association rate ($K_g^{TF_1}$) (Fig. 3B). When we generate rates for the rules, the concentration (C^{TF_1}) is accounted for by the SSA because the firing propensity of any rule considers the number of molecules available. The SRB applies a similar process to the nucleosome abstraction using the nucleosome affinity score [21] to define sequence preferences.

Each rule generated by the SRB creates a set of reactants and resultants, which define the possible states of each nucleotide. Each nucleotide can only be bound by a single factor at a given time, which is a temporal constraint in our framework that is based on biophysical constraints. The SRB must ensure that each bound state of the nucleotide is mutually exclusive by augmenting the state names with the bound factor name (see supplemental for details). This is conceptually similar to each nucleotide having a limited number of possible states in the HMM formalisms (Fig S1).

The transcriptional machinery is capable of loading onto DNA and starting transcription at any position. However, it also traverses DNA and terminates transcription at given rates. We have set the default rates for transcribing from the in vitro experimental data [76]. In addition, there are cases where specific transcription factors can bind to DNA and recruit the transcriptional machinery to a nearby position [77]. For example, TATA binding protein (TBP) recruits the transcriptional machinery to a DNA position roughly 35 bases downstream of its position. Therefore, the SRB allows each transcription factor to recruit the transcriptional machinery to a position at a specified distance (shown in appendix Fig. A6).

An interesting modeling issue arises when the transcriptional machinery encounters obstacles (other components) during its movement. Given the stochastic nature of factor binding, the transcriptional machinery could just wait until the obstacle removes itself. However, it is more likely that the transcriptional machinery actively removes obstacles [78]. Consequently, we allow the transcriptional machinery to modify the eviction probabilities of a protein when it encounters an obstacle. When the obstacle is another transcriptional machine traversing in the opposite direction, this results in transcriptional interference [66], [79]. For simplicity, in our current implementation, transcription will abort in both directions when interference occurs.

3.4.2 Coping with parameters

An inherent obstacle to our approach is the large number of necessary parameters (Tables S4-S8). Whenever possible, we obtain the default parameters from the literature. For example, our modeling of the yeast genome uses protein counts obtained from Ghaemmaghami [80]. The PSSM for transcription factors were obtained from MacIsaac [81] and Badis [33]. Nucleosome affinity is from Kaplan [82] and Wasson [21]. In all of these cases, the datasets were generated from population-averaged experiments, typically with cells grown in standard yeast media (YPD) and measured during log phase growth. Therefore, they are considered baseline default values that can be overridden by the user in a modeling configuration file. User specified custom motifs are also supported, allowing newly discovered proteins to be quickly incorporated.

The temporal events also require rates. In general, these rates are largely unknown and currently difficult to estimate from the literature. While recent advances in experimental approaches [44] show tremendous promise in rapidly addressing this parametrization issue, currently many of the kinetic parameters have not been experimentally determined. Some of the factor 'on' rates are known, but very few 'off' rates (or residency times) are known [11]. We estimate

these rates by tying them to other well-studied parameters. Experimental studies indicate that the temporal information is unlikely to be fully independent of the positional information, as the sequence being bound influences the kinetics of the reaction [11], [83]. Therefore, we assume that higher affinity sites will also have higher residency times [11](Fig. S9). Hence, a site with strong sequence preference will bind more frequently and for a longer period of time. These rates can be overridden by user defined global rates or data derived position dependent information provided in the configuration file.

3.4.3 The system overview

The SRB generates potentially thousands of chemical reaction rules. These can then be fed to off-the-shelf simulators (Fig. 2). In our case, the SRB output is formatted for the third-party stochastic simulation engine Dizzy [84]. Briefly, we use the Next Reaction Method [85], which is an extension of an exact stochastic simulation algorithm [63]. In these simulations, the state of the system is represented by the set of current molecule counts of each reactant and the time until each of the reactions will occur. The times are selected from the probability distributions for each reaction, which are based on the rate of the reaction and the current molecular counts of the reactants. After initialization of the internal data structures, the algorithm repeatedly selects the next reaction to occur and applies that reaction. Each reaction changes the molecule counts, which may affect the probability distribution of many other reactions using the same reactants. The algorithm recalculates, as needed, the time to next reaction for all the affected reactions. As many independent reactions could occur nearly simultaneously, many reactions may be applied during each time step. Time, as defined by Dizzy, is an arbitrary unit in which a number of actions occur. At the end of each time step the current molecule counts for all reactants are reported. The algorithm continues to apply reactions until the user specified number of time steps has been reached.

In our models, each possible state of a nucleotide becomes a possible reactant. The size of our models therefore depends on the size of the DNA segment being modeled and the number of different protein components being used. Because each action influences only a localized span of nucleotides, the worst case is always a linear growth in the number of rules (Fig. 4). When all possible rules are applied at every position, this growth is still substantial (Fig. S6).

The number of rules produced can be reduced by the application of biologically realistic and reasonable heuristics. For example, while in theory a transcription factor can bind to any location, it is anticipated that significant interactions will only occur with sequences with reasonably good fit to the PSSM [86]. Applying a score cutoff to the PSSM reduces the number of rules significantly, but makes the number of rules strongly dependent on the underlying sequence. We also optionally allow rules to be applied to groups of nucleotides, a heuristic that drastically reduces the number of rules produced for a specified length of DNA, but also reduces the precision of the simulation. Our default system typically generates hundreds of thousands of interaction rules for a typical gene and millions of rules for chromosomes (Table S9).

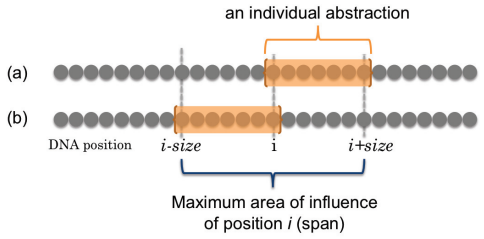

Fig. 4. Each DNA position influences a limited number of rules. Focusing at the nucleotide level, the model captures the fact that each nucleotide (grey circle) can be in only one of a limited number of states. (a) Each abstraction (orange box) has a span of nucleotides that are influenced when the Petri net is applied at position i. (b) Therefore, a chance at any position i, influences a fixed window size $[i - size, i + size]$ of nearby nucleotides, defined by the span of the largest abstraction.

There are practicalities with using a simulation method that must also be considered. First, we initialize all the DNA states as unbound and allow the system to populate binding factors until a quasi-equilibrium is reached. Through observation, we have found 500 steps to be sufficient for this burn-in period. Each simulation is run for a minimum of 8000 steps, to allow for sampling of a variety of pathways. Finally, because we frequently simulate small sections of DNA, there are edge effects. The behavior of the model is altered at the edges of the model (beginning and end of the DNA region) due to an absence of defined rules for positions outside the modeled region. The natural barrier of the edges of the region can be felt for a distance along the DNA. In our experience, padding the edges by two thousand nucleotides, a distance in excess of three times the largest rule span, reduces edge effects.

The stochastic simulation engine, Dizzy, requires the rules and reactants description file as input, as well as specification of the method, number of time steps, and the sampling rate. Each simulation results in a different series of events and describes one possible trajectory of the given DNA through time. Dizzy's implementation is a Monte Carlo implementation and scales logarithmically with the number of rules [84]. While we have chosen Dizzy because of its availability and relative ease of use, the SRB output could easily be revised to use any off-the-shelf SSA simulator (FERN [87], NFsim [88], DYNSTOC [89]).

Finally, the output results of the simulations can be summarized or visualized. Running the simulation multiple times or for an extended period of time will stochastically select a different sequence of events, leading to alternative trajectories. Combining many time points allows for summary statistics, such as the distribution of binding configurations, to be obtained and compared to experimental data. In addition, the simulation results can be interpreted in light of the possible actions described by our Petri nets, allowing for the visualization of the configuration of each position at each time step, as well as inferring the dynamic movement of factors along the DNA (Fig. S4).

4 Case Studies

In this section we describe case studies that verify the simulations capture both the positional and temporal aspects of regulation at well-studied yeast loci. Each case study is

focused on validating a distinct property of our framework. The first study examines our ability to capture positional information in a manner similar to the best positional models. The second case study focuses on the ability to capture temporal information, such as the effects of transcriptional interference as a regulatory mechanism. Finally, we consider the scalability of our framework by modeling an entire chromosome. The end goal of these simulations is to provide validation of our modeling framework, not to present new biological insight. Here we describe the overall results of these case studies and interesting observations arising from our work. A description of the parameters, runtimes, memory usage, and the outputs of each simulation are included in the Supplemental Material (Fig. S11-S13).

To evaluate each of the models, we compared the model's predicted nucleosome occupancy with an experimentally measured data set [90]. While our model is at base pair resolution, the Lee experimental data is measured at a 4 base pair resolution. Therefore, we summarize our predicted values over each Lee data probe (+/- 1 bp) and compare the two using a Pearson correlation.

4.1 Capturing positional information (GAL/CLN2)

The first case study focuses the model's ability to capture accurate positional information. Occupancy is a population averaged measure of the time that a position of the DNA is by a given protein. Changes in transcription factor concentration can alter the identity and occupancy of factors within the region. For example, the GAL locus is one of the most well studied regions within the yeast genome. This locus has a known activator (GAL4), which has multiple binding locations in the promoter region. By running our simulation at distinct concentrations, our model shows a parallel increase in the occupancy of GAL4, causing shifts in nucleosome positioning that open the chromatin near both GAL10 and GAL1 transcription start sites (Fig. S11). Our model and the best positional models, as exemplified by the Wasson model (COMPETE) [21], both predict the changes in nucleosome occupancy in the region (using GAL4 and nucleosomes only).

A more interesting case is the CLN2 locus, where experiments show that multiple factors work cooperatively, without explicit protein-protein interaction, to maintain a known nucleosome depleted region (NDR) [5]. Bai and colleagues determined that the NDR of CLN2 was dependent on three transcription factors: Mcm1, Reb1, and Rsc3. We sought to confirm that our modeling framework could capture this nucleosome depleted region in a manner similar to the state-of-the-art positional models, again using COMPETE (Fig. 5). In this case, we quantitatively compared each model to Lee's experimentally determined nucleosome occupancy profile, obtained at 4-nucleotide resolution [90]. Giving each model (COMPETE and our framework) the three key transcription factors (Mcm1, Reb1, and Rsc3), both models correlated well with the Lee data (COMPETE r=0.52; our SSA model r=0.51$\pm$.07) showing that, similar to GAL4, our framework captures the same steady state behaviors as a state of the art positional model (COMPETE).

We next sought to understand the underlying dynamics implied by our simulations between the transcription factors, the nucleosomes, and the DNA sequence at this locus. When our SSA model uses only nucleosomes, the correlation to the experimental data is little better than random (r=0.10)(data not shown). This result is consistent with previous experimental work which showed the NDR had lower nucleosome occupancy than was predicted by positional models that use only nucleosome affinity [82]. Next, we add in only the transcription factor Mcm1 and observe the emergence of a well-positioned nucleosome at one edge of the NDR. This nucleosome influences the positioning of adjacent nucleosomes, consistent with studies that indicate that a single transcription factor can impact many nucleosomes [91]. As we add additional transcription factors to the model, they improve the correlation with experimental data, showing that our framework can capture the implicit cooperation of multiple binding factors to maintain the NDR.

Finally, we sought to determine how sensitive our model results were to the particular choices made for the less well determined parameters within our model, namely binding rates and molecule counts. We found that small changes in molecule counts for Mcm1 have an initial dramatic effect on configurations, but that saturation is quickly reached (data not shown). Our stable NDR results were obtained using three transcription factors (Mcm1, Rsc3, Reb1) at the default molecule counts [80] by adjusting the OFF rate, which controls the residency time. If the OFF rate is reduced to 10%, the molecule count must be increased by a factor of 10 to maintain the NDR. This implies a balancing act between molecule counts and residency time that is consistent with intuition. Furthermore, it has been speculated that modulation of residency times may allow for more precise regulation [11].

It should be noted that the single trajectory nature of our simulations leads to a number of potentially interesting observations at the CLN2 locus. First, all three transcription factors are known to be required to maintain the NDR [5], but our models show that they are not necessarily required to be bound simultaneously. We observe time points where zero, one, two or three transcription factors are bound, limiting the nucleosome positions and preserving the NDR. In addition, the NDR shows some limited binding of nucleosomes, consistent with the fact that a depleted region does not imply the absence of binding, but rather less binding than expected. The result of each simulation provides an independent prediction for nucleosome occupancies and the correlation with the experimental data yields a range of values having a median of 0.53 and a maximum of 0.63 (Fig. S12). Lastly, we observe that a transcription factor with longer residency time has a stronger effect on nearby nucleosome positioning (data not shown). This effect propagates out from the transcription factor position over time, leading to speculation that longer residency times may be necessary for long range effects. While these results may reflect phenomena that arise from our specific parametrization, we found these general patterns to be fairly robust. Furthermore, our model allows for the exploration of the temporal patterns of interactions that lead to the NDR result, observations that could be experimentally validated.

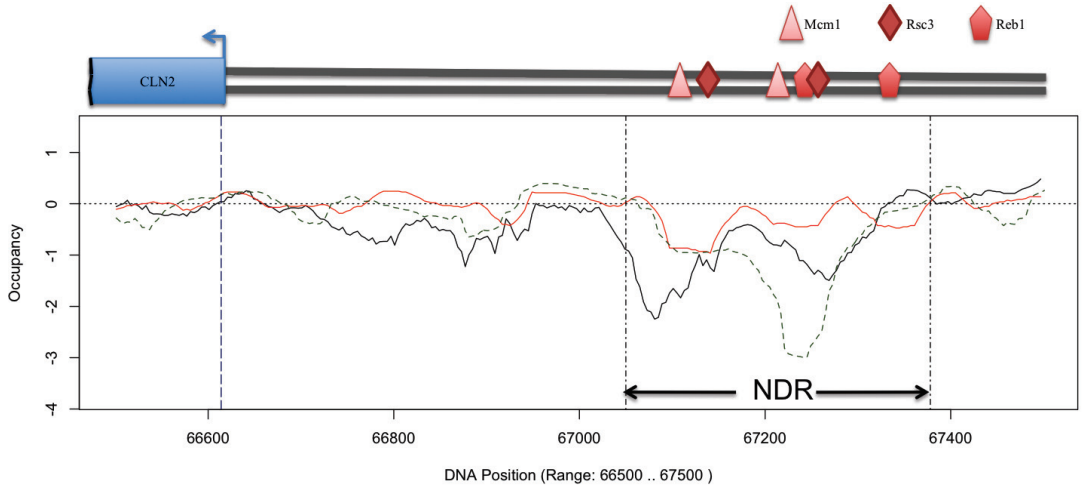

Fig. 5. Modeling nucleosome occupancy at gene CLN2 recapitulates the known nucleosome depleted region (NDR). The yeast gene CLN2 (blue box) contains a well-studied nucleosome depleted region (NDR) upstream of the start site. The known location of three key transcription factors (Mcm1, Rsc3, and Reb1) are shown as red geometric shapes [5]. The experimentally determined nucleosome occupancy [90] is shown in black, as observed from ChIP signal (normalized, log2 scale). The results from our model (red line) and the COMPETE model (green dotted line), plotted as the log ratio of a predicted nucleosome occupancy (y-axis) as a function of DNA position (x-axis). Graph depicts *S. cerevisiae* chromosome XVI positions 64000-72000.

4.2 Capturing temporal interaction (IME4)

The second case study focuses on the ability of our framework to capture temporally driven events, such as transcriptional interference. We use a well-studied location, the IME4 locus, where transcription of the gene is constitutive and largely unregulated. Yet, Ime4 is an N6-adenosine methyltransferase required only for a cell's entry into meiosis. To modulate Ime4 levels, the cell produces an antisense transcript constitutively that, by transcriptional interference, stops the gene's sense transcript from completing [46]. Only in diploids where the antisense transcript is itself repressed can the full length transcript of IME4 be produced. It is precisely this sort of regulation that motivated the development of our modeling system.

We sought to confirm that our model could recapitulate the known transcriptional interference pattern observed in experimental data. In yeast there exists a transcription factor, a/α, which is unique to diploids. We ran the IME4 locus both with and without the a/α transcription factor. In the absence of the transcription factor, we observe robust transcription of the antisense transcript and very little of the sense transcript reaches full length. In the presence of a/α, the antisense transcript is repressed by occluding the antisense initiation site, allowing most of the sense transcript to reach full length. These simulation results are consistent with the experimental data at IME4 [46], [92]. We then explored a series of interesting "what if" scenarios at this locus. For instance, we studied the impact of different transcriptional machinery initiation and elongation rates on the transcriptional interference event. From these parameter explorations, it is clear that the initiation rate of the transcriptional machinery must be fast enough to always keep at least one polymerase transcribing in the antisense direction for each initiation of sense strand machinery (Fig. S13). This pattern is consistent with previous mathematical studies of transcriptional interference [66].

Another observation is worth noting, transcription through any transcription start site causes a transient localized open chromatin conformation that makes initiating transcriptional machinery more likely. Thus, the few sense transcripts that do complete, i.e., the transcriptional machinery traverses through the start location of the antisense transcript, typically trigger the initiation of an antisense transcript. This may illustrate a particularly interesting feedback mechanism where the rate of transcription through a region regulates the open chromatin and therefore transcription at nearby sites.

4.3 Tractability

Our last case examines the tractability of using our framework to model large systems. One of our goals was to include temporal information while maintaining computability for large sequences. This case study specifically focuses on testing options for scaling the simulations. We chose to model an entire chromosome from the yeast genome; *S. cerevisiae* chromosome I is approximately 230,000 nucleotides, with 92 genes.

In the worst case, modeling of a large sequence with many transcription factors can generate millions of rules and overwhelm the simulation engines. When using transcription factor threshold cutoffs, a single locus, such as GAL, CLN2, or IME4, generates thousands of rules and reactants (Table S9). For chromosome I, a similar approach at single nucleotide resolution would produce approximately 7 million reactants and 40 million rules. To make this simulation tractable, we employed nucleotide grouping, reducing the granularity of the resultant simulation. We found that grouping the DNA into 30 nucleotide units reduced the model to 170,000 reactants and 400,000 rules. At this size, Dizzy was tractable, requiring 98 GB of ram and running for just over 20 CPU hours per simulation.

Our resource limiting factor in computability is the stochastic simulation engine, currently Dizzy [84]. It com-

putes large tables to efficiently transition between the possible interactions. As these tables grow towards a given machine's available RAM, performance decreases quickly. Managing the number of rules that are created keeps the models computable. It is possible that other off-the-shelf simulators (e.g. FERN [87], NFsim [88], DYNSTOC [89]) would be capable of larger simulations using less compute resources. Alternatively, it is possible to replace the off-the-shelf stochastic simulation engine with one designed specifically for this application.

5 Discussion

Our goal was to integrate inherently dynamic aspects of transcriptional regulation, such as transcriptional interference, with the intuitive position based approach in order to model the eukaryotic transcription process in a tractable, sequence specific fashion. To this end, we constructed a modeling framework that leverages the power of Petri nets to describe the actions of various regulators and the extent or span of their influence. By treating the DNA as an ordered set of entities (nucleotides or groups of nucleotides) rather than a single molecular entity, we can generate models that grow linearly with the length of the DNA sequence being modeled. At the core of our framework is our stochastic rule builder, an application that can take in an arbitrary sequence and construct the complete set of coherent biochemical rules. Off-the-shelf stochastic simulation engines, such as Dizzy, can then simulate these rule sets.

We have developed a framework to create biologically realistic models of the mechanisms of transcriptional regulation. Based on this framework, we can model not only the steady-state behavior of transcription factor binding and nucleosome formation (case study 1), but also the dynamics of components, such as the transcriptional machinery (case study 2). Our framework scales linearly, making it possible to simulate very large segments of DNA (case study 3). The simulations produce tremendous amounts of positional and temporal data, which can be converted into simple visualizations depicting the state of the DNA at each time step (see Fig. S4).

There is an intimate relationship between the development of new experimental techniques and new modeling frameworks. Typically, the level of detail of the modeling abstraction is influenced by both the questions being asked and the resolution of available experimental data. Large-scale experimental techniques are continuously evolving to capture increasingly detailed views of regulation. Recent experimental work is only beginning to highlight the importance of temporal dynamic events, such as transcription factor turnover [11], nucleosome turnover [93], and transcriptional interference [7], for understanding transcriptional regulation.

As is true with any new modeling system, our framework depends on a number of parameters to capture the distinct behaviors of individual components. We require not only the DNA binding parameters used in positional models, but also rate parameters that capture the underlying temporal aspects of events. Unfortunately, many of the temporal parameters are not currently known. We presently set these parameters using coarse searches for values that reasonably capture desired phenomena or fit available experimental data. We are well aware that many of these parameters may be over-fit, thus detailed parameter explorations and sensitivity analysis remains as future work. However, as new experimental studies uncover these rates or identify new key regulatory mechanisms, our modeling framework is poised to incorporate this information. Ultimately, single cell measurements [14] will permit a more precise comparison between our model and biological reality. In the meantime, we envision our framework as an exploratory tool that allows modelers to rapidly prototype loci with different parameters or DNA sequences.

As our understanding of the molecular details improves, our framework can be easily extended. Currently, our system simplifies every component in an attempt to capture the essence of its behavior. However, it may be necessary to extend existing components to capture key molecular events. For example, we currently consider the nucleosome as a single large binding component. Yet in reality, the nucleosome is composed of a histone core (H3-H4), which binds first, and two subunits (H2A-H2B) on each edge [94]. Recent work on nucleosome dynamics indicates the core histone is relatively stable, whereas the edge histones are more dynamic [95], [96]. In the future, we may need to model the core and edge components separately to account for their differences in behavior. Likewise, we may need to add additional components, such as histone tail modifications, that can affect the affinity or stability of a nucleosome. While histone modifications are known to be well correlated with transcriptional state [97], the details of how these marks influence binding, when these marks are temporally deposited, and how they function is not well characterized. Therefore, the addition of new components or augmented functionality must be balanced against the parametrization problem. Only as we understand the dynamic behavior of these components in more detail is it realistic to include these within our framework.

Our framework results in a large set of rules to describe the chemical reactions within the system. We currently utilize Dizzy as our underlying stochastic simulation engine. There are many other engines available [87], [88], [89], some of which may permit larger models or shorter run times to be obtained. An alternative to simulating the system of equations is to mathematically solve them. Solving the system of differential equations would provide the equilibrium behavior of the system, but requires large-scale system solvers. Even the best of these solvers are limited in their ability to handle tens of thousands of equations [98]. Simulation can handle much larger sets of equations, but at the cost of increasing computational time.

Finally, our SRB produces single trajectory simulations for a single cell, but they are not whole cell simulations. The current models reach a relative equilibrium (cycling through common configurations), as there is little feedback to alter protein concentrations or modify component behaviors within the system. As the parameters of the model become more informed by experimental data, we envision introducing realistic feedback into the model.

6 Conclusion

We have created a modeling framework that captures both the positional and temporal aspects of transcriptional regulation. Our framework uses Petri nets to describe permissible actions and their localized span of influence. We have created an application, the stochastic rule builder, which quickly generates large systems of chemical reactions to model any specific instance of DNA. The resulting equations can be simulated with standard stochastic simulation engines. We confirmed through case studies that our model can capture positional information, temporal information, and is scalable to large segments of DNA. We consider our framework, at this time, as an early proof of principle eukaryotic framework. As technological advances in single cell experimentation further uncover temporal cellular kinetics, our models will continue to advance towards biologically realistic and predictive models of transcriptional regulation.

Acknowledgments

We appreciate the comments and suggestions from anonymous reviewers that improved the clarity of this paper. This work is supported by the Boettcher Foundation's Webb Waring Biomedical Research Program and NSF Grant ABI 1262410. DAK was also supported by a National Science Foundation under Grant DGE 084142 (eCSite) and by the NLM Informatics Training Grant 2T15 LM009451-07.

References

[1] P. J. Farnham, "Insights from genomic profiling of transcription factors." *Nature reviews. Genetics*, vol. 10, no. 9, pp. 605–616, Sep. 2009.

[2] A. H. Mack, D. J. Schlingman, R. P. Ilagan, L. Regan, and S. G. J. Mochrie, "Kinetics and thermodynamics of phenotype: unwinding and rewinding the nucleosome." *Journal of molecular biology*, vol. 423, no. 5, pp. 687–701, Nov. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22944905

[3] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer, "In vivo dynamics of RNA polymerase II transcription." *Nature structural & molecular biology*, vol. 14, no. 9, pp. 796–806, Sep. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17676063

[4] L. A. Mirny, "Nucleosome-mediated cooperativity between transcription factors." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 52, pp. 22 534–22 539, Dec. 2010. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0913805107

[5] L. Bai, A. Ondracka, and F. R. Cross, "Multiple Sequence-Specific Factors Generate the Nucleosome-Depleted Region on CLN2 Promoter." *Molecular cell*, vol. 42, no. 4, pp. 465–76, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.molcel.2011.03.028

[6] S. Hahn and E. T. Young, "Transcriptional regulation in saccharomyces cerevisiae: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators," *Genetics*, vol. 189, no. 3, pp. 705–736, Nov. 2011.

[7] A. C. Palmer, J. B. Egan, and K. E. Shearwin, "Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors." *Transcription*, vol. 2, no. 1, pp. 9–14, Jan. 2011.

[8] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thå ström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom, "A genomic code for nucleosome positioning." *Nature*, vol. 442, no. 7104, pp. 772–8, Aug. 2006.

[9] B. J. Venters, S. Wachi, T. N. Mavrich, B. E. Andersen, P. Jena, A. J. Sinnamon, P. Jain, N. S. Rolleri, C. Jiang, C. Hemeryck-Walsh, and B. F. Pugh, "A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces." *Molecular cell*, vol. 41, no. 4, pp. 480–92, Feb. 2011.

[10] R. K. Bradley, X.-Y. Li, C. Trapnell, S. Davidson, L. Pachter, H. C. Chu, L. a. Tonkin, M. D. Biggin, and M. B. Eisen, "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species." *PLoS biology*, vol. 8, no. 3, p. e1000343, Mar. 2010.

[11] C. R. Lickwar, F. Mueller, S. E. Hanlon, J. G. McNally, and J. D. Lieb, "Genome-wide proteinDNA binding dynamics suggest a molecular clutch for transcription factor function," *Nature*, vol. 484, no. 7393, pp. 251–255, Apr. 2012.

[12] J. M. Levsky, S. M. Shenoy, R. C. Pezo, and R. H. Singer, "Single-cell gene expression profiling." *Science (New York, N.Y.)*, vol. 297, no. 5582, pp. 836–40, Aug. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12161654

[13] E. a. Galburt, S. W. Grill, and C. Bustamante, "Single molecule transcription elongation," *Methods*, vol. 48, no. 4, pp. 323–332, Aug. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19426807

[14] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells." *Science (New York, N.Y.)*, vol. 329, no. 5991, pp. 533–538, Jul. 2010.

[15] D. R. Larson, D. Zenklusen, B. Wu, J. a. Chao, and R. H. Singer, "Real-time observation of transcription initiation and elongation on an endogenous yeast gene." *Science (New York, N.Y.)*, vol. 332, no. 6028, pp. 475–478, Apr. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21512033

[16] E. Segal and J. Widom, "From DNA sequence to transcriptional behaviour: a quantitative approach." *Nature reviews. Genetics*, vol. 10, no. 7, pp. 443–56, Jul. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19506578

[17] K. Struhl and E. Segal, "Determinants of nucleosome positioning." *Nature structural & molecular biology*, vol. 20, no. 3, pp. 267–73, Mar. 2013. [Online]. Available: http://www.nature.com/doifinder/10.1038/nsmb.2506

[18] A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev, "Effect of promoter architecture on the cell-to-cell variability in gene expression." *PLoS computational biology*, vol. 7, no. 3, p. e1001100, Mar. 2011.

[19] T.-L. To and N. Maheshri, "Noise can induce bimodality in positive transcriptional feedback loops without bistability." *Science (New York, N.Y.)*, vol. 327, no. 5969, pp. 1142–5, Feb. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20185727

[20] D. Zeevi, E. Sharon, M. Lotan-Pompan, Y. Lubling, Z. Shipony, T. Raveh-Sadka, L. Keren, M. Levo, A. Weinberger, and E. Segal, "Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters." *Genome research*, vol. 21, no. 12, pp. 2114–28, Dec. 2011.

[21] T. Wasson and A. J. Hartemink, "An ensemble model of competitive multi-factor binding of the genome." *Genome research*, vol. 19, no. 11, pp. 2101–12, Nov. 2009.

[22] A. D. Lander, "The edges of understanding." *BMC biology*, vol. 8, p. 40, Jan. 2010.

[23] H. Kim and E. Gelenbe, "Stochastic gene expression modeling with Hill function for switch-like gene responses." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 9, no. 4, pp. 973–9, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22144531

[24] C. P. Barnes, D. Silk, X. Sheng, and M. P. H. Stumpf, "Bayesian design of synthetic biological systems." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 37, pp. 15 190–5, Sep. 2011.

[25] I. Cantone, L. Marucci, F. Iorio, M. A. Ricci, V. Belcastro, M. Bansal, S. Santini, M. di Bernardo, D. di Bernardo, and M. P. Cosma, "A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches." *Cell*, vol. 137, no. 1, pp. 172–81, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19327819

[26] A. S. Ribeiro, O.-P. Smolander, T. Rajala, A. Häkkinen, and O. Yli-Harja, "Delayed stochastic model of transcription at the single nucleotide level." *Journal of computational biology*, vol. 16, no. 4, pp. 539–53, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19361326

[27] S. Lubliner and E. Segal, "Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy." *Bioinformatics (Oxford, England)*, vol. 25, no. 12, pp. i348–55, Jun. 2009.

[28] S. J. Greive, J. P. Goodarzi, S. E. Weitzel, and P. H. von Hippel, "Development of a "modular" scheme to describe the kinetics of transcript elongation by RNA polymerase." *Biophysical journal*, vol. 101, no. 5, pp. 1155–65, Sep. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21889453

[29] M. R. Roussel and R. Zhu, "Stochastic kinetics description of a simple transcription model." *Bulletin of mathematical biology*, vol. 68, no. 7, pp. 1681–713, Oct. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16967259

[30] M. Levo and E. Segal, "In pursuit of design principles of regulatory sequences." *Nature reviews. Genetics*, vol. 15, no. 7, pp. 453–68, Jul. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24913666

[31] B. Lemon and R. Tjian, "Orchestrated response: A symphony of transcription factors for gene control," *Genes and Development*, vol. 14, no. 20, pp. 2551–2569, Oct. 2000. [Online]. Available: http://www.genesdev.org/cgi/doi/10.1101/gad.831000

[32] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK21054/

[33] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-f. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, "Diversity and complexity in DNA recognition by transcription factors." *Science (New York, N.Y.)*, vol. 324, no. 5935, pp. 1720–3, Jun. 2009.

[34] C. Zhu, K. J. R. P. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk, "High-resolution DNA-binding specificity analysis of yeast transcription factors." *Genome research*, vol. 19, no. 4, pp. 556–66, Apr. 2009.

[35] H. R. Drew and a. a. Travers, "DNA bending and its relation to nucleosome positioning." *Journal of molecular biology*, vol. 186, no. 4, pp. 773–90, Dec. 1985. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/3912515

[36] A. V. Morozov, K. Fortney, D. a. Gaykalova, V. M. Studitsky, J. Widom, and E. D. Siggia, "Using DNA mechanics to predict in vitro nucleosome positions and formation energies." *Nucleic acids research*, vol. 37, no. 14, pp. 4707–22, Aug. 2009.

[37] P. T. Lowary and J. Widom, "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning." *Journal of molecular biology*, vol. 276, no. 1, pp. 19–42, Feb. 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9514715

[38] H. Sung Rhee and B. Franklin Pugh, "Genome-wide structure and organization of eukaryotic pre-initiation complexes," *Nature*, vol. 487, no. 7389, pp. 128–128, Mar. 2012.

[39] K. D. MacIsaac and E. Fraenkel, "Practical strategies for discovering regulatory DNA sequence motifs," *PLoS Computational Biology*, vol. 2, no. 4, pp. 201–210, Apr. 2006.

[40] A. M. Sengupta, M. Djordjevic, and B. I. Shraiman, "Specificity and robustness in transcription control networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 4, pp. 2072–7, Feb. 2002.

[41] O. G. Berg, R. B. Winter, and P. H. von Hippel, "Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory." *Biochemistry*, vol. 20, no. 24, pp. 6929–48, Nov. 1981. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7317363

[42] J. L. Workman, "Nucleosome displacement in transcription." *Genes & development*, vol. 20, no. 15, pp. 2009–17, Aug. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16882978

[43] O. I. Kulaeva, F.-K. Hsieh, H.-W. Chang, D. S. Luse, and V. M. Studitsky, "Mechanism of transcription through a nucleosome by RNA polymerase II." *Biochimica et biophysica acta*, vol. 1829, no. 1, pp. 76–83, Jan. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22982194

[44] A. Coulon, C. C. Chow, R. H. Singer, and D. R. Larson, "Eukaryotic transcriptional dynamics: from single molecules to cell populations." *Nature reviews. Genetics*, vol. 14, no. 8, pp. 572–84, Aug. 2013. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg3484

[45] E. M. Prescott and N. J. Proudfoot, "Transcriptional collision between convergent genes in budding yeast." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 13, pp. 8796–8801, Jun. 2002.

[46] C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink, "Antisense transcription controls cell fate in Saccharomyces cerevisiae." *Cell*, vol. 127, no. 4, pp. 735–45, Nov. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17110333

[47] S. L. Bumgarner, R. D. Dowell, P. Grisafi, D. K. Gifford, and G. R. Fink, "Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 43, pp. 18 321–6, Oct. 2009.

[48] P. Thebault, G. Boutin, W. Bhat, A. Rufiange, J. Martens, and A. Nourani, "Transcription Regulation by the Noncoding RNA SRG1 Requires Spt2-Dependent Chromatin Deposition in the Wake of RNA Polymerase II." *Molecular and cellular biology*, vol. 31, no. 6, pp. 1288–300, Mar. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21220514

[49] S. J. Hainer, J. A. Pruneski, R. D. Mitchell, R. M. Monteverde, and J. A. Martens, "Intergenic transcription causes repression by directing nucleosome assembly." *Genes & development*, vol. 25, no. 1, pp. 29–40, Jan. 2011.

[50] V. Pelechano, S. Chávez, and J. E. Pérez-Ortín, "A complete set of nascent transcription rates for yeast genes." *PloS one*, vol. 5, no. 11, p. e15442, Jan. 2010.

[51] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell." *Science (New York, N.Y.)*, vol. 297, no. 5584, pp. 1183–6, Aug. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12183631

[52] B. B. Kaufmann and A. van Oudenaarden, "Stochastic gene expression: from single molecules to the proteome." *Current opinion in genetics & development*, vol. 17, no. 2, pp. 107–12, Apr. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17317149

[53] S. Huang, "Non-genetic heterogeneity of cells in development: more than just noise." *Development*, vol. 136, no. 23, pp. 3853–62, Dec. 2009.

[54] A. Schwabe, M. Dobrzyski, K. Rybakova, P. Verschure, and F. J. Bruggeman, "Origins of stochastic intracellular processes and consequences for cell-to-cell variability and cellular survival strategies." *Methods in enzymology*, vol. 500, pp. 597–625, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21943916

[55] N. Tenazinha and S. Vinga, "A survey on methods for modeling and analyzing integrated biological networks." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 8, no. 4, pp. 943–58, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21116043

[56] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks." *Nature reviews. Molecular cell biology*, vol. 9, no. 10, pp. 770–80, Oct. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18797474

[57] A. Ay and D. N. Arnosti, "Mathematical modeling of gene expression: a guide for the perplexed biologist." *Critical reviews in biochemistry and molecular biology*, vol. 46, no. 2, pp. 137–51, Apr. 2011.

[58] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models-a review." *Bio Systems*, vol. 96, no. 1, pp. 86–103, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19150482

[59] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules." *Nature structural biology*, vol. 9, no. 9, pp. 646–52, Sep. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12198485

[60] A. S. Ribeiro, "Stochastic and delayed stochastic models of gene expression and regulation." *Mathematical biosciences*, vol. 223, no. 1, pp. 1–11, Jan. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19883665

[61] C. Chaouiya, "Petri net modelling of biological networks." *Briefings in bioinformatics*, vol. 8, no. 4, pp. 210–9, Jul. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17626066

[62] A. Sanchez, S. Choubey, and J. Kondev, "Stochastic models of transcription: From single molecules to single cells," *Methods*, vol. 62, no. 1, pp. 13–25, Jul. 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23557991

[63] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled

chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, Dec. 1976. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0021999176900413

[64] A. J. Black and A. J. McKane, "Stochastic formulation of ecological models and their applications." *Trends in ecology & evolution*, vol. 27, no. 6, pp. 337–45, Jun. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22406194

[65] K. Takahashi, K. Kaizu, B. Hu, and M. Tomita, "A multi-algorithm, multi-timescale method for cell simulation." *Bioinformatics (Oxford, England)*, vol. 20, no. 4, pp. 538–46, Mar. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14990450

[66] K. Sneppen, I. B. Dodd, K. E. Shearwin, A. C. Palmer, R. A. Schubert, B. P. Callen, and J. B. Egan, "A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli." *Journal of molecular biology*, vol. 346, no. 2, pp. 399–409, Feb. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15670592

[67] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences." *Nucleic acids research*, vol. 18, no. 20, pp. 6097–100, Oct. 1990.

[68] M. L. Kireeva, B. Hancock, G. H. Cremona, W. Walter, V. M. Studitsky, and M. Kashlev, "Nature of the nucleosomal barrier to RNA polymerase II." *Molecular cell*, vol. 18, no. 1, pp. 97–108, Apr. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15808512

[69] J. Mäkelä, J. Lloyd-Price, O. Yli-Harja, and A. S. Ribeiro, "Stochastic sequence-level model of coupled transcription and translation in prokaryotes." *BMC bioinformatics*, vol. 12, no. 1, p. 121, Jan. 2011.

[70] R. Zhu, A. S. Ribeiro, D. Salahub, and S. a. Kauffman, "Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models." *Journal of theoretical biology*, vol. 246, no. 4, pp. 725–45, Jun. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17350653

[71] M. J. Schilstra and C. L. Nehaniv, "Stochastic model of template-directed elongation processes in biology." *Bio Systems*, vol. 102, no. 1, pp. 55–60, Oct. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20655359

[72] H. Genrich, R. Kuffner, and K. Voss, "Executable Petri net models for the analysis of metabolic pathways," *Int J STTT*, vol. 3, pp. 394–404, 2001.

[73] D. Ruths, M. Muller, J.-T. Tseng, L. Nakhleh, and P. T. Ram, "The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks." *PLoS computational biology*, vol. 4, no. 2, p. e1000005, Feb. 2008.

[74] I. Mura and A. Csikász-Nagy, "Stochastic Petri Net extension of a yeast cell cycle model." *Journal of theoretical biology*, vol. 254, no. 4, pp. 850–60, Oct. 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18703074

[75] J. Lei, "A Modular, Qualitative Modeling of Regulatory Networks Using Petri Nets," *Modeling in Systems Biology*, vol. 16, p. 25, 2011. [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-84996-474-6_12

[76] S. F. Tolić-Nø rrelykke, A. M. Engh, R. Landick, and J. Gelles, "Diversity in the rates of transcript elongation by single RNA polymerase molecules." *The Journal of biological chemistry*, vol. 279, no. 5, pp. 3292–9, Jan. 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14604986

[77] G. O. Bryant and M. Ptashne, "Independent recruitment in vivo by Gal4 of two complexes required for transcription." *Molecular cell*, vol. 11, no. 5, pp. 1301–9, May 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12769853

[78] M. A. Schwabish and K. Struhl, "Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II." *Molecular and cellular biology*, vol. 24, no. 23, pp. 10 111–7, Dec. 2004.

[79] M. Gullerova and N. J. Proudfoot, "Transcriptional interference and gene orientation in yeast: noncoding RNA connections." *Cold Spring Harbor symposia on quantitative biology*, vol. 75, pp. 299–311, Jan. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21467144

[80] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, "Global analysis of protein expression in yeast." *Nature*, vol. 425, no. 6959, pp. 737–41, Oct. 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14562106

[81] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel, "An improved map of conserved regulatory sites for Saccharomyces cerevisiae." *BMC bioinformatics*, vol. 7, p. 113, Jan. 2006.

[82] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, "The DNA-encoded nucleosome organization of a eukaryotic genome." *Nature*, vol. 458, no. 7236, pp. 362–366, Mar. 2009.

[83] J. D. Lieb, X. Liu, D. Botstein, and P. O. Brown, "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." *Nature genetics*, vol. 28, no. 4, pp. 327–34, Aug. 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11455386

[84] S. Ramsey, D. Orrell, and H. Bolouri, "Dizzy: stochastic simulation of large-scale genetic regulatory networks." *Journal of bioinformatics and computational biology*, vol. 3, no. 2, pp. 415–36, May 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15852513

[85] M. A. Gibson and J. Bruck, "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels," *The Journal of Physical Chemistry A*, vol. 104, no. 9, pp. 1876–1889, Mar. 2000. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/jp993732q

[86] Y. Qi, A. Rolfe, K. D. MacIsaac, G. K. Gerber, D. Pokholok, J. Zeitlinger, T. Danford, R. D. Dowell, E. Fraenkel, T. S. Jaakkola, R. a. Young, and D. K. Gifford, "High-resolution computational models of genome binding events." *Nature biotechnology*, vol. 24, no. 8, pp. 963–70, Aug. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16900145

[87] F. Erhard, C. C. Friedel, and R. Zimmer, "FERN - a Java framework for stochastic simulation and evaluation of reaction networks." *BMC bioinformatics*, vol. 9, p. 356, Jan. 2008.

[88] M. W. Sneddon, J. R. Faeder, and T. Emonet, "Efficient modeling, simulation and coarse-graining of biological complexity with NFsim." *Nature methods*, vol. 8, no. 2, pp. 177–183, 2011.

[89] J. Colvin, M. I. Monine, J. R. Faeder, W. S. Hlavacek, D. D. Von Hoff, and R. G. Posner, "Simulation of large-scale rule-based models." *Bioinformatics (Oxford, England)*, vol. 25, no. 7, pp. 910–7, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19213740

[90] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow, "A high-resolution atlas of nucleosome occupancy in yeast." *Nature genetics*, vol. 39, no. 10, pp. 1235–44, Oct. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17873876

[91] A. Jansen, E. van der Zande, W. Meert, G. R. Fink, and K. J. Verstrepen, "Distal chromatin structure influences local nucleosome positions and gene expression." *Nucleic acids research*, vol. 40, no. 9, pp. 3870–85, May 2012.

[92] B. Gelfand, J. Mead, A. Bruning, N. Apostolopoulos, V. Tadigotla, V. Nagaraj, A. M. Sengupta, and A. K. Vershon, "Regulated antisense transcription controls expression of cell-type-specific genes in yeast." *Molecular and cellular biology*, vol. 31, no. 8, pp. 1701–1709, Apr. 2011.

[93] M. F. Dion, T. Kaplan, M. Kim, S. Buratowski, N. Friedman, and O. J. Rando, "Dynamics of replication-independent histone turnover in budding yeast." *Science (New York, N.Y.)*, vol. 315, no. 5817, pp. 1405–8, Mar. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17347438

[94] A. J. Andrews and K. Luger, "Nucleosome structure(s) and stability: variations on a theme." *Annual review of biophysics*, vol. 40, pp. 99–117, Jan. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21332355

[95] G. Li, M. Levitus, C. Bustamante, and J. Widom, "Rapid spontaneous accessibility of nucleosomal DNA." *Nature structural & molecular biology*, vol. 12, no. 1, pp. 46–53, Jan. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15580276

[96] V. Böhm, A. R. Hieb, A. J. Andrews, A. Gansen, A. Rocker, K. Tóth, K. Luger, and J. Langowski, "Nucleosome accessibility governed by the dimer/tetramer interface." *Nucleic acids research*, vol. 39, no. 8, pp. 3093–102, Apr. 2011.

[97] S. L. Berger, "Histone modifications in transcriptional regulation," *Current Opinion in Genetics & Development*, pp. 142–148, 2002.

[98] J. Pahle, "Biochemical simulations: stochastic, approximate stochastic and hybrid approaches." *Briefings in bioinformatics*, vol. 10, no. 1, pp. 53–64, Jan. 2009.

David A. Knox David A. Knox received a BA in 1982 from the University of Colorado Boulder. After 35 years in the computer industry, he returned to Boulder to obtain a Masters in CS in 2010 and is working toward his PhD in Computational Bioscience at the University of Colorado Anschutz Medical Campus. David has been a member of IEEE since 1981.

Robin D. Dowell Robin D. Dowell received two BS degrees (Genetics, Computer Engineering) in 1997 from Texas A&M University, a Masters in Computer Science from Washington University in St. Louis in 2001, and a D.Sc in Biomedical Engineering from Washington University in St. Louis in 2004. She is now an Assistant Professor at the University of Colorado in the BioFrontiers Institute and the Molecular, Cellular and Developmental Biology Department. Robin has been a member of IEEE since 2001.