Parametric modeling of protein-DNA binding kinetics: A discrete event based simulation approach

Preetam Ghosh, Samik Ghosh, Kalyan Basu and Sajal Das

Biological Networks Research Group, Dept. of Computer Science & Engineering
University of Texas at Arlington, Arlington, TX-76010.
Email: {ghosh, sghosh, basu, das}@cse.uta.edu

Abstract

To understand the stochastic behavior of biological systems, we adopt an “in silico” stochastic event based simulation methodology that can determine the temporal dynamics of different molecules. The main requirement for this technique is the event execution time models for the different biological functions of the system. This paper presents a parametric model to determine the execution time of one such biological function: protein-DNA binding. This biological function is modeled as a combination of microlevel biological events using a coarse grained probability measure to estimate the stochastic parameters of the function. Our model considers the actual binding mechanism along with some approximated protein and DNA structural information. We use a collision theory based approach to transform the thermal and concentration gradients of this biological function into their corresponding probability measure. This information theoretic approach significantly removes the complexity of the classical protein sliding along the DNA model, improves the speed of computation and can bypass the speed-stability paradox. This model can produce acceptable estimates of DNA-protein binding time necessary for our event-based stochastic system simulator where the higher order (more than second order statistics) uncertainties can be ignored. The results show good correspondence with available experimental estimates. The model depends very little on experimentally generated rate constants and brings the important biological parameters and functions into consideration. We also present some “in silico” results showing the effects of protein-DNA binding on gene expression in prokaryotic cells.

Key words: Stochastic Modeling, Biophysics, Systems Biology, Collision Theory.

Preprint submitted to Elsevier 25 November 2007
1 Introduction

The system simulation of biological processes is important to understand their dynamics. Recent molecular level measurements of biological processes have identified a stochastic resonance [1,2] specially for protein creation and other signaling pathways. The stochastic simulation models [3,4,5,6] using the approximate master equation are based on rate equations. Due to the large number of proteins in a cell, these models lead to combinatorial explosion in the number of reactions, thus making them unmanageable for complex signaling pathway problems. Our goal is to build a stochastic discrete event based framework [7] for biological systems to overcome the computational complexity of current mesoscale and stochastic simulation methods. This flexible simulation framework has the potential to be extended to a genome scale simulation.

We consider a biological system as a collection of biological processes, each comprising a number of functions, and a function is modeled as an event with relevant boundary conditions. These events are threaded by the pathway logic to develop a stochastic discrete-event simulation. For the simulation, we need to identify the main parameters that impact these biological events. The events are modeled using the abstraction of the biological function as a series of microevents. The measure of the uncertainty of the microevents is used to create the stochastic behavior of the event and the statistics are obtained by using applied probability theory. The description of the simulation method can be found in [7] and the abstraction mechanisms in [8,9,10,11]. Here, we extend the event modeling approach to compute the execution time of another complex biological function: ‘DNA-protein’ binding.

The transcription factors (TFs), bind DNA at specific sites to initiate the complex transcription machinery of cells. Upon binding to the site, the TF forms a stable protein-DNA complex that can either activate or repress transcription of nearby genes, depending on the actual control mechanism. In this paper, we focus on models for both bacterial and eukaryotic TFs and we assume that the structure, location on chromatin and other details of cognate (target) sites on the DNA can be approximated from existing experimental data. Such problems of specific binding and binding rates arise in the context of oligonucleotides-DNA binding [12] as well. In the current model, we have not included the effects of chromatin remodeling and histone modifications.

2 Related Works

Vast amounts of experimental data available these days provide the structures of protein-DNA complexes at atomic resolution in crystals and in solution [13,14,15], binding constants for dozens of native and hundreds of mutated proteins [16,17], calorimetry measurements [18] and novel single-molecule ex-
periments [19]. Based on these experimental data, a conceptual basis for describing both the kinetics and thermodynamics of protein-DNA interactions was first presented in [20,21,22,23]. The classical model of protein-DNA sliding however, is quite complicated. The problem that the sliding mechanism faces if the energetics of protein-DNA interactions are taken into account is outlined in [24], where the authors introduce a quantitative formalism for protein-DNA interactions.

2.1 Previous model - Protein sliding along the DNA

The existing TF-DNA binding model involves a combination of both three-dimensional (3-d) and one-dimensional diffusion (1-d) of the TF. The total search process can be considered as a 3-d search followed by binding to the DNA and a round of 1-d diffusion. The TF, upon dissociation from the DNA, continues on a 3-d diffusion until it binds at a different place on the DNA. The 1-d diffusion along the DNA proceeds along the rough energy landscape of the DNA. A quantitative analysis of the search process in [24] reported the following:

1. The diffusion along the DNA becomes prohibitively slow when the roughness of the binding energy landscape is at least $2k_B T$.
2. The optimal energy of nonspecific binding to the DNA provides the maximal search rate. However, even the optimal combination of 1-d and 3-d diffusions cannot achieve experimental estimates of binding time when the roughness of the landscape is at least $2k_B T$. In the optimal regime of search, the protein spends equal amounts of time diffusing along non-specific DNA (i.e., 1-d diffusion) and diffusing in the solution (i.e., 3-d diffusion). A fairly smooth landscape (with roughness of the order of $k_B T$) is required for the 1-d diffusion to achieve experimentally observed and biologically relevant rates.
3. Stability of the protein-DNA complex at the target site requires considerably larger roughness than $k_B T$ where rapid search is impossible, leading to the speed-stability paradox. In fact, the minimal roughness as reported in [24] is $5k_B T$ given a genome size of $10^6$ bps. A search-and-fold mechanism for the DNA-binding proteins is proposed in [24] to resolve the paradox.

While the TF diffusion along the DNA is controlled by the specific binding energy (i.e., energy required for the TF to bind to a particular DNA sequence), the dissociation of the TF from the DNA depends on the total binding energy (i.e., on the non-specific binding as well as on the specific one). Moreover, since the dissociation events are much less frequent than the hopping between neighboring base-pairs, the non-specific energy makes a larger contribution to the total binding energy.
2.2 Our Contributions

We consider the binding for both bacterial and eukaryotic transcription factors (TFs) to the DNA assuming that the structure, location on chromatin and other details of target sites on the DNA are known. This data can be found from the existing biological databases (e.g., [25,26]) or need to be determined from experiments if they are not. In contrast to the existing thermodynamic and diffusion based models, our approach closely follows the biological process that involves a number of discrete microevents. We assume that the TF binding site of the DNA is exposed with a difficulty factor depending upon the location of the site with respect to the nucleosome.

The main idea is that for bacterial cells, the TF (with matching motif) randomly collides with the DNA and only when it hits the binding site with enough kinetic energy to overcome the energy barrier of the site, can the binding occur. Based on our research focus, we abstract the first micro biological event ‘collision of the TF to the DNA surface’ by using the collision theory model for non-spherical collision objects. The information measure we compute from this abstraction is the probability of DNA-protein collision. The next microlevel biological event is the binding of a TF to the DNA based on the description of the protein and DNA structures on the chromatin as encountered in the biological process.

This method bypasses the speed-stability paradox of protein-DNA interactions to allow for a computationally efficient model. Our stochastic simulator uses this fast model in a similar way as the rate constants used by the Gillespie simulator [3] to approximate the protein-DNA binding time. The TF sliding mechanism due to thermal gradient, for searching the binding region is also incorporated in our model and we show that not all DNA-TF collisions result in sliding. For eukaryotic cells, the protein-DNA binding mechanism is achieved in two steps 1) diffusion of the TF to the nucleus of the cell and 2) random collisions of the TF with the DNA (we assume that the TF never comes out of the nucleus) for the binding. Our model computes the entire DNA-protein binding time for bacterial cells and DNA-protein binding time once the protein has entered the nucleus for eukaryotic cells. The average time for diffusion of protein molecules to the nucleus can be easily computed from standard diffusion models. We validate our model for the DNA replication process in prokaryotic cells. We also present some “in silico” results showing the effects of protein-DNA binding on gene expression in prokaryotic cells.

3 DNA-Protein Binding Model

We partition this problem into two biological microevents: (i) Collision of the protein molecule to a binding site ($\pm B$) on the DNA surface, i.e., we assume
that the TF can slide a distance of $B$ (in either direction) on the DNA before binding and (ii) a protein colliding with DNA at the binding site ($\pm B$) will bind only if it hits it with enough kinetic energy to overcome the energy barrier of the site.

3.1 Modeling the first microevent: calculating $p_n$

In this section we abstract the first microevent by computing the probability, $p_n$, of collision of the protein (TF) with the binding site ($\pm B$) on the DNA. From the principles of collision theory for hard spheres, we model the protein molecule as a rigid sphere with diameter $d$ and the TF binding region of the DNA as a solid cylinder with diameter $D$ and length $L + 2B$ (Fig 1). Note that the $2B$ factor is incorporated as the TF can slide in either direction on the DNA.

![Fig. 1. Schematic diagram of protein molecule and TF binding region of the DNA: $D$ and $L + 2B$ are the diameter and length of the TF-binding region, and $d$ is the diameter of the TF.](image)

We define our coordinate system such that the DNA is stationary with respect to the protein molecule. This assumption allows the TF to move towards the DNA with relative velocity $U$. The protein molecule moves through space to sweep out a collision cross section, $C$. The number of collisions during a time period $\Delta t$ is determined when a protein molecule will be inside the space created by the motion of the collision cross section over this time period due to the motion of the protein molecule.

3.1.1 Calculating the average surface area of collision between a sphere and cylinder

The spherical protein molecule during its motion can encounter the DNA binding sites in three different configurations (1) horizontal cylinder, (2) vertical...
Fig. 2. Collision of spherical protein and cylindrical DNA TF binding region.

cylinder and (3) cylinder at an arbitrary angle, θ, with the direction of motion of the protein (Fig 2). For the horizontal cylinder model, the cross-sectional area of collision traces out a circle, whereas for the vertical cylinder model, it is cylindrical in shape. The third case can be derived from the vertical cylinder model considering a cylindrical collision area of length \((L + 2B + d)\sin\theta\). Thus, the cross-sectional area of collision, \(C\), is given by:

\[
C = \begin{cases} 
\pi \frac{(d + D)^2}{4}, & \text{for } \theta = 0^0 \\
(L + 2B + d)(D + d), & \text{for } \theta = 90^0 \\
(D + d)(L + 2B + d)\sin\theta, & \text{otherwise} 
\end{cases}
\]

Thus for any arbitrary θ \((0^0 < \theta < 90^0)\), we can express the cross-sectional area of collision as a function of θ as follows: \(C(\theta) = (D + d)(L + 2B + d)\sin\theta\).

Note that the border conditions \((\theta = 0^0, 90^0)\) constitute a set of measure zero and for all practical purposes, the whole calculation can be limited to the case where \(0^0 < \theta < 90^0\). We assume a uniform density for the occurrence of the different θ’s in the range \(0^0 \leq \theta \leq 90^0\), i.e. having density \(\frac{\theta}{(\pi/2)}\). It is to be noted that ideally θ can take any value in \(0^0 \leq \theta \leq 360^0\), but our working range of \(0^0...90^0\) suffices for all these cases. Thus the average cross-sectional area, \(C_{avg}\), can be expressed by:
\[ C_{\text{avg}} = \int_{0}^{\frac{\pi}{2}} \frac{2}{\pi} C(\theta) d\theta = \frac{2}{\pi} (D + d)(L + 2B + d). \]

Note that \( \theta = 0^\circ \) disappears from consideration but we can argue that the probability of that happening is too small to change the expression for \( C_{\text{avg}} \) significantly. This cross-section \( C_{\text{avg}} \), moves in the cytoplasmic space (nucleus for eukaryotes) to create the collision volume for a particular binding site.

### 3.1.2 Probability of protein-DNA binding in eukaryotic cells

Fig 3 simplistically illustrates how DNA is packed along different cylindrical nucleosomes. We do not include chromatin remodeling and histone modification in the current model as discussed in Section 5.6. Thus, \( L \) in the expression for \( C_{\text{avg}} \) denotes the length of the TF binding region and \( D \) the diameter of the DNA strand (assumed cylindrical in shape) on a nucleosome cylinder. As single or multiple motifs [27] can be present for a gene in the promoter region, the value of \( L \) is adjusted to reflect those conditions. Now, we can have three cases based on where the TF binding region is located on the DNA:

1) Case I: The region entirely lies within the DNA portion on a nucleosome cylinder;
2) Case II: The region lies entirely within the DNA portion that is outside the nucleosome cylinders;
3) Case III: The region is shared between the DNA on a nucleosome cylinder and that outside it.

We analyze each of these cases in the following:

*Case I:* Let the probability that the protein molecule hits the correct nucleosome cylinder given it collided with the DNA with sufficient energy be \( p_0^c \). We
have:

\[ p_h^c = \frac{\text{length of that nucleosome cylinder}}{\text{length of all nucleosomes} + \text{length of all stretches}} \]

\[ = \frac{l_n}{N_1 l_n + \sum_{i=1}^{N_2} l_i^s} \]

where, \( l_n \) denotes the length of a nucleosome cylinder (assumed fixed for all the cylinders), \( l_i^s \) denotes the length of the \( i^{th} \) stretch of DNA, i.e., the length of DNA present in between the \( i^{th} \) and \((i + 1)^{th}\) nucleosome cylinders. Here, \( N_1 \) and \( N_2 \) denote the number of nucleosome cylinders and that of stretches of DNA respectively. Now, the probability, \( p_d \), of hitting the DNA portion of the nucleosome cylinder, can be estimated from the surface area of the nucleosome cylinder and that of the DNA present in the cylinder as follows:

\[ p_d = \frac{\pi Dl_d}{\pi Dl_d + \pi d_n l_n} \]

where, \( l_d \) is the length of the DNA present inside the cylinder and \( d_n \) is the diameter of the nucleosome cylinder. Because the DNA is known to make 1.65 turns in a nucleosome cylinder, we have \( \frac{l_d}{l_n} = 1.65 \). Let, \( p_f^c \) designate the probability of colliding with the TF binding region (±B) in the DNA, given that the protein molecule already collided with the DNA with enough energy and also hit the correct nucleosome cylinder. We have:

\[ p_f^c = \frac{(\text{length of TF binding region in the DNA}) + 2B}{\text{total DNA length in that particular nucleosome}} \]

Also, the particular motif of the colliding protein molecule is of interest to us, as it should come in proximity to the TF binding region (±B) of the DNA for a binding to occur. So, we need to calculate the probability, \( p_m \), of identifying the motif of the colliding protein molecule, as follows:

\[ p_m = \frac{\text{length of the motif region of the protein}}{\text{total length of amino acid chain of the protein}} \]

Thus, the total probability of collision of the TF to the DNA binding site (±B) is given by:

\[ p_n = p_m \times p_h^c \times p_f^c \times p_d \]

Now, because the DNA is wrapped around a particular nucleosome cylinder, some part of it will not be available for the TF to bind to. Thus \( C_{avg} \) as
calculated above is not entirely available to the TF to bind to. Nucleosomes themselves are stable and show limited mobility. The dynamic characteristics are due to the actions of the nucleosome-modifying and remodeling complexes that restructure, mobilize and eject nucleosomes to regulate access to the DNA. We approximate the impact of this complex process currently through a difficulty parameter $\alpha$, which denotes the \textit{percentage availability in average collision cross-sectional area}. This parameter represents approximately the percentage of the time the hidden DNA surface is made visible for reaction through histone remodeling (we are currently working on a separate model of histone remodeling to compute this parameter). Thus, the effective cross-sectional area, $C_{eff}$, available for TF binding can be calculated as follows: $C_{eff} = \alpha \times C_{avg}$.

\textbf{Case II:} In this case, the probability, $p_h^s$, of hitting the correct stretch of DNA in between the nucleosome cylinders is given by:

$$p_h^s = \frac{l_i^i}{N_1 l_n + \sum_{i=1}^{N_2} l_i^i}$$

where we assume that the TF binding site is located in the $i^{th}$ stretch of DNA. Similarly, let $p_f^s$ designate the probability of colliding with the TF binding region ($\pm B$) in the DNA similarly as before. We have:

$$p_f^s = \frac{(\text{Length of TF binding region on DNA}) + 2B}{\text{total DNA length in that particular stretch}}$$

and, the total probability of collision of the TF to the DNA binding site denoted by $p_n$ is given by:

$$p_n = p_m \times p_h^s \times p_f^s$$

In this case, the entire TF binding region in the DNA is available for the binding process to occur and we have: $C_{eff} = C_{avg}$.

\textbf{Case III:} Because the TF binding region ($\pm B$) is shared between a nucleosome cylinder and an adjoining stretch, the probability calculations become complex for this case. We approximate the calculations in the following way. Suppose the TF binding site ($\pm B$) is shared between the $i^{th}$ nucleosome cylinder and the $j^{th}$ stretch of DNA. Because the cylinder and the stretch has to be side by side, we must have either $j = i$, or $i = j + 1$ depending on whether the first part of the TF binding site is in the cylinder or in the stretch respectively. Let $p_w$ and $p_w^s$ denote the probabilities of hitting the TF binding portion in the cylinder and that in the stretch respectively. In this case however, $p_f^s$ and $p_f^j$ computations should change as follows:
\[ p^*_f = \frac{\text{(length of TF binding region portion in nucleosome)} + B}{\text{total length of DNA in that particular nucleosome}} \]
\[ p^*_s = \frac{\text{(length of TF binding region portion in the stretch)} + B}{\text{total length of DNA in that particular stretch}} \]

Hence we have:

\[ p^*_w = p_m \times p^*_h \times p^*_f \times p_d; \quad p^*_w = p_m \times p^*_h \times p^*_f; \quad p_n = p^*_w + p^*_w \]

Thus \( p_n \) is the total probability of collision of the TF to the DNA binding site (\( \pm B \)). Furthermore, the average cross-sectional area calculations become a little different in this case. We break up \( C_{avg} \) into \( C_{avg1} \) and \( C_{avg2} \) based on the length of the TF binding region (\( L_1 \)) in the nucleosome cylinder and that in the adjoining stretch (\( L_2 \)). We assume for simplicity that the TF binding region is shared between one stretch and one nucleosome cylinder only, because this region is generally quite small in length compared to the length of the DNA packed inside a nucleosome cylinder. However, if the region is extended to more than one nucleosome cylinder or stretch, we can handle that case in a similar fashion. Thus the effective cross-sectional area of binding is represented as:

\[ C_{eff} = \alpha \times C_{avg1} + C_{avg2} \]

Thus the total probability, \( p_n \), of collision to one specific TF binding region can be calculated easily for each of the three cases discussed above. But we need to know how exactly the DNA is packed in the nucleosome cylinders to determine \( p_n \) and the effective surface area (\( C_{eff} \)) required for binding. In particular, we assume that the DNA packing structure in nucleosome cylinders is fixed and hence we can find where the TF binding region is located as described in Cases I, II and III.

### 3.1.3 Approximate mechanism to find the TF binding region

Nucleosomes have 1.65 turns of DNA and a diameter, \( d_n \), of 11 nm. Thus the length of DNA inside a nucleosome cylinder can be approximated as \( 1.65 \times \pi \times d_n \), where \( \pi d_n \) is the circumference of the nucleosome cylinder. We assume that all the nucleosome cylinders have identical shape and number of turns of DNA in them. We also assume that all the stretches of DNA between nucleosome cylinders are equal in length. Thus, length of DNA in a stretch can be approximated as \( \frac{T_D - N \times (1.65 \times \pi \times d_n)}{N-1} \), where \( T_D \) is the total length of the DNA and \( N \) is the number of nucleosome cylinders present. The denominator is due to the assumption that there can only be \( N - 1 \) stretches of DNA present in between the \( N \) nucleosome cylinders. From the complete genomic
sequence, we can find out the exact position of the TF binding region along with its length. Thus we can approximately estimate whether the TF binding region corresponds to Case I, II or III.

3.1.4 Protein-DNA binding probability for bacterial cells

The bacterial genome is supercoiled with a general organization as depicted in Fig 4 [28]. Each domain consists of a loop of DNA, the ends of which are secured in some way. Hence, the total probability of collision in this case is simply approximated as:

\[ p_n = p_m \times p_w; \text{ where } p_w = \frac{\text{length of TF binding region} + 2B}{\text{total length of the DNA}} \]

Since the entire surface area of the DNA is available for binding, the effective cross-sectional area of binding is given by: \( C_{eff} = C_{avg} \)

3.2 Modeling the second microevent: calculating \( p_b \)

Let \( p_b \) denote the probability that the TF collides with the DNA with enough kinetic energy such that it can bind to the DNA. In time \( \Delta t \), the TF sweeps out a volume \( \Delta V \) such that:

\[ \Delta V = C_{eff} U \Delta t \]

Now, the probability of the protein molecule being present in the collision volume \( \Delta V \) is \( p_P = 1 \) given that one protein molecule arrived to create a collision volume of \( \Delta V \).
The probability of the DNA being present in an arbitrary uniformly distributed $\Delta V$ in the total volume, $V$ is given by $p_D = \frac{\Delta V}{V}$. Note that the prokaryotic cells do not have a nucleus and hence $V$ denotes the total volume of the cell; for eukaryotic cells, however, $V$ will denote the volume of the nucleus.

Thus, probability of the protein molecule to collide with the DNA during time $\Delta t$ is:

$$p_c = p_P \times p_D = \frac{\Delta V}{V} = \frac{C_{eff} U \Delta t}{V}$$

(1)

We next assume that the colliding protein molecule must have free energy of at least $E_{Act}$ to bind to the specific DNA transcription factor binding region. This kinetic energy will be required for the rotational motion of the protein molecule such that all the binding points in the protein molecule come close to those in the DNA for the binding to take place successfully. The kinetic energy of approach of the protein towards the DNA with a velocity $U$ is $E = \frac{m_{PD} U^2}{2}$, where $m_{PD} = \frac{m_P m_D}{m_P + m_D}$ is the reduced mass, $m_P$ is the mass (in gm) of the protein molecule, and $m_D$ is the mass (in gm) of the DNA. It is to be noted that we consider the mass of the entire chromosome and not just the TF binding site of the DNA. This is because the entire chromosome has to undergo rotational motion for the binding process. We also assume that as the kinetic energy, $E$, linearly increases above $E_{Act}$, the number of collisions that result in binding also increases. Thus, the probability for a binding to occur because of sufficient kinetic energy of the protein molecule is given by:

$$p_r = \begin{cases} \frac{E - E_{Act}}{E}, & \text{for } E > E_{Act} \\ 0, & \text{otherwise} \end{cases}$$

(2)

and the overall probability, $p_o$, for collision with sufficient energy is given by:

$$p_o = p(binding, Collision) = p_r \times p_c = \begin{cases} p_c \frac{(E - E_{Act})}{E}, & \text{for } E > E_{Act} \\ 0, & \text{otherwise}. \end{cases}$$

The above equations assume a fixed relative velocity $U$ for the reaction. We will use the Maxwell-Boltzmann distribution of molecular velocities for a species of mass $m$ given by:

$$f(U, T)dU = 4\pi \left( \frac{m}{2\pi k_B T} \right)^{3/2} e^{-\frac{mU^2}{2k_BT}} U^2 dU$$
where \( k_B = 1.381 \times 10^{-23} \, \text{kg} \, \text{m}^2/\text{s}^2/\text{K/molecule} \) is the Boltzmann’s constant and \( T \) denotes the absolute temperature (taken as 273° K). Replacing \( m \) with the reduced mass \( m_{PD} \) of the protein molecule and DNA, we get

\[
f(U, T)dU = 4\pi \left( \frac{m_{PD}^2}{2\pi k_B T} \right)^{3/2}e^{-\frac{m_{PD} U^2}{2k_BT}} U^2 dU
\]

The term on the left hand side of the above equation denotes the fraction of this specific protein molecule with relative velocities between \( U \) and \((U + dU)\). Summing up the collisions for the protein molecule for all velocities, the probability \( (p_b) \) of collision with sufficient energy is obtained as follows:

\[
p_b = \int_0^\infty p_o f(U, T)dU
\]

Recalling that \( E = \frac{m_{PD} U^2}{2} \), i.e., \( dE = m_{PD} UdU \) and substituting into Eqn. 3, we get:

\[
f(U, T)dU = 4\pi \left( \frac{m_{PD}^2}{2\pi k_B T} \right)^{3/2} \frac{2E}{U m_{PD}^2} e^{-\frac{E}{k_BT}} dE
\]

Thus we get:

\[
p_b = \int_{E_{Act}}^\infty \frac{(E - E_{Act}) 4C_{eff} \Delta t}{V k_B T} \sqrt{\frac{1}{2\pi k_B T m_{PD}}} e^{-\frac{E}{k_BT}} dE
\]

\[
= \frac{C_{eff} \Delta t}{V} \sqrt{\frac{8k_B T}{\pi m_{PD}}} \frac{e^{-\frac{E_{Act}}{k_BT}}}{e^{-\frac{E}{k_BT}}}
\]

### 3.3 Total binding probability considering different binding regions

Ideally, for any protein molecule, we can have more than one TF binding regions on the DNA. Let \( G \) be the number of different TF binding regions on the DNA for the specific TF colliding with the DNA. Also, let \( p_i^t \) denote the total probability of binding (combining the first and second microevents) for the \( i^{th} \) TF binding region \((1 \leq i \leq G)\). Note that the probabilities of the first and second microevents as calculated above will depend on the specific binding site \( i \) on the DNA under consideration. We denote these two probabilities as
$p^i_n$ and $p^i_b$ for the $i^{th}$ site that can be calculated similarly as shown above. In general, all the binding sites corresponding to a particular TF are identical making $p^i_n = p^j_n$ and $p^i_b = p^j_b$, for $i \neq j$, and $1 \leq i, j \leq G$. Hence,

$$p^i_t = p^i_n \times p^i_b$$

Thus if $p$ denotes the actual probability of binding of the protein with any of these $G$ different regions, we have:

$$p = \sum_{i=1}^{G} [p^i_t \prod_{j=1, j \neq i}^{G} (1 - p^j_t)]$$

This is because the probability of binding to the first TF binding region is given by $p^1_t \prod_{j=2}^{G} (1 - p^j_t)$; that for the second region is $[p^2_t (1 - p^1_t)(1 - p^2_t) ...(1 - p^G_t)]$; and so on. The total probability, $p$, is the sum of all these individual cases.

4 Time taken for protein-DNA binding

We next estimate the time taken to complete the binding with total binding probability, $p$. Let $\Delta t = \tau$ be an infinitely small time step. The protein molecules try to bind to the DNA through collisions. If the first collision fails to produce a successful binding, they collide again after $\tau$ time units and so on. Note that now we can have a TF-DNA binding in two ways: (a) the TF directly collides and binds to the DNA binding site or (b) the TF collides at a distance ($\leq B$ bps) and slides on the DNA to bind to the site. The average binding time computation requires a probability assignment to these two events. Let $per$ denote the probability that the binding occurs due to collision only (point (a) above). Hence, binding occurs with collision and sliding with probability $(1 - per)$. Note that $per = 1$ simplifies to the case where the protein does not slide along the DNA at all, and $per = 0$ boils down to the model in [24] where it is assumed that the TF slides along the DNA at every round. In [24], the authors derived the 1-d diffusion time, $\tau_{1d}$, along the DNA using the mean first passage time (MFPT) from site 0 to $B$ as follows:

$$\tau_{1d}(B) \simeq B^2 e^{\frac{\sigma^2}{4(k_B T)^2}} (\nu)^{-\frac{1}{2}} (1 + \frac{\sigma^2}{2(k_B T)^2})^{-\frac{1}{2}}$$

where $\nu$ is the effective attempt frequency for hopping to a neighboring site and $\sigma$ is the roughness of the DNA landscape in units of $k_B T$. Here $\tau_{1d}$ considers the different energy barriers on the DNA that the TF has to overcome while sliding
whereas $E_{act}$ is required for the actual binding to the cognate site. Therefore, the total probability of binding is:

$$p_{\text{binding}} = p_{\text{no-sliding}}(1 - p) + p(1 - p_{\text{no-sliding}}); \text{ and, } p_{\text{no-sliding}} = |p|_{B=0}$$

where $p_{\text{no-sliding}}$ denotes the probability of binding when the sliding along the DNA is not considered altogether. Hence, the average time for protein-DNA binding model (i.e., the first moment) is given by:

$$T_1 = p_{\text{binding}}(\text{per} \times \tau + (1 - \text{per})(\tau + \tau_{ld})) + (1 - p_{\text{binding}})p_{\text{binding}} \times 2(\text{per} \times \tau + (1 - \text{per})(\tau + \tau_{ld})) + (1 - p_{\text{binding}})^2p_{\text{binding}} \times 3(\text{per} \times \tau + (1 - \text{per})(\tau + \tau_{ld})) + ...$$

$$\Rightarrow T_1 = \frac{(\text{per} \times \tau + (1 - \text{per})(\tau + \tau_{ld}))}{p_{\text{binding}}}$$

The second moment of the binding time is given by

$$T_2 = \frac{(2 - p_{\text{binding}})(\text{per} \times \tau + (1 - \text{per})(\tau + \tau_{ld}))^2}{(p_{\text{binding}})^2}$$

When no sliding is considered, we find that the time for DNA-protein binding follows an exponential distribution for most ranges of $E_{act}$ (reported in the next section). Moreover, since $\tau$ is assumed to be quite small, we can approximate the total time measurements of binding using a continuous (exponential in this case) distribution instead of a discrete geometric distribution. The average time $T_1$ as calculated above gives the estimated time for protein-DNA binding in bacterial cells. For eukaryotic cells, we should add the average protein transport time from the cytoplasm to the nucleus that can be computed from any standard diffusion model.

5 Results and analysis

5.1 Problems in validation of our model

Before presenting the numerical results, let us first discuss the difficulty of experimentally validating our model. We compute the average time for protein-DNA binding in this chapter. On the other hand, existing experimental results are based on estimation of the binding rate of any specific TF to the DNA.
The experimental estimate of $1 \sim 10$ seconds (secs) is reported from this rate measurement [24] for the PurR TF in E. coli. Hence, the time taken by a TF to bind to the DNA site depends on the number of TFs in the cell. However, our model computes the time taken by any particular TF to bind to the DNA which should be independent of the number of TFs in the cell. It is certainly very difficult to carry out experiments to track a particular TF and physically compute the time. Also, the stochastic nature of the binding process suggests that the distribution of the time taken will have a very high variance. In other words, in some cases the TF requires time in milliseconds whereas in other cases it might take as long as 100 seconds. The results we present in this section assume that the time taken for the PurR TF-DNA binding is $1 \sim 10$ secs even though it is not a true estimate of this event because it is not a molecular level measurement.

5.2 Numerical results for \( \text{per} = 1 \) (i.e. no TF sliding is considered)

In this section, we present the numerical results for the theoretical models derived in the chapter. Figs 5-8 present the results for the PurR TF (having 35 binding sites) on the Escherichia coli (\(E. \ coli\)) chromosome. Similarly, Figs 9-10 illustrate the behavior for eukaryotic cells where we considered the average human cell with 20 µm diameter and the Htrf1 DNA-binding protein. The different parameters assumed for the numerical results are concisely presented in Table 1. We used the EcoCyc database [25] for the \(E. \ coli\) data and the PDB database [26] for human cell data.

5.2.1 Results for prokaryotic cells

Fig 5 plots \(T_1\) against different values for \(\Delta t\). The average time for DNA-protein binding remains constant initially and shoots up exponentially with increasing \(\Delta t\). The same characteristics are seen for different activation energies, \(E_{\text{act}} = 10 \ k_B T, 15 \ k_B T\) and \(20 \ k_B T\). The activation energy estimates follow from the change in free energy related to binding that includes the entropic loss of translational and rotational degrees of freedom of the protein and amino acid side chains, the entropic cost of water and ion extrusion from the DNA surface, the hydrophobic effect, etc. as discussed in [29]. The smaller the required \(E_{\text{act}}\), the larger is \(p_b\) for the protein molecules and hence the smaller is \(T_1\). Note that \(p_b\) as calculated above also corresponds to the number of collisions in time \(\Delta t\) of the protein molecule with the DNA. And, for our assumption of at most one collision taking place in \(\Delta t\) to hold, we have to make sure that \(0 \leq p_b \leq 1\) (this is also true because \(p_b\) is a probability). Thus the regions to the right of the vertical lines corresponding to each \(E_{\text{act}}\) plot denotes the forbidden region where \(p_b > 1\) even though \(0 \leq p \leq 1\). This
Table 1
Parameter Estimation for Bacterial (pertaining to PurR TF in *E. coli*) and Eukaryotic (pertaining to Htrf1 TF in human) Cells

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Prokaryotic Cell (from [25])</th>
<th>Eukaryotic Cell (from [26])</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>V</em> (volume)</td>
<td>$4.52 \times 10^{-18} \text{m}^3$ (of cell)</td>
<td>$4.187 \times 10^{-16} \text{m}^3$ (of nucleus)</td>
</tr>
<tr>
<td>Length of DNA</td>
<td>$4.64 \times 10^6 \text{bp}$ (<em>E. coli</em>)</td>
<td>$3 \times 10^9 \text{bp}$ (<em>Human cell</em>)</td>
</tr>
<tr>
<td><em>G</em> (number of binding regions)</td>
<td>35 (for PurR)</td>
<td>35 (assumed for Htrf1)</td>
</tr>
<tr>
<td>Length of TF binding site (<em>L</em>)</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td>Length of protein amino acid chain</td>
<td>341 (for PurR)</td>
<td>53 (Htrf1)</td>
</tr>
<tr>
<td>Length of protein motif</td>
<td>26 (for PurR)</td>
<td>48 (Htrf1)</td>
</tr>
<tr>
<td>Radius of Amino acid chain</td>
<td>1 nm (for PurR)</td>
<td>1 nm (Htrf1)</td>
</tr>
<tr>
<td>Average radius of the protein (<em>$\bar{d}$</em>)</td>
<td>5 Å (for PurR)</td>
<td>5 Å (Htrf1)</td>
</tr>
<tr>
<td>$m_p$</td>
<td>38.175 Dalton (for PurR)</td>
<td>6635 Dalton (for Htrf1)</td>
</tr>
<tr>
<td>Diameter of DNA (<em>D</em>)</td>
<td>2 nm (for <em>E. coli</em>)</td>
<td>2 nm (<em>Human cell</em>)</td>
</tr>
<tr>
<td>Mass of DNA (<em>$m_D$</em>)</td>
<td>$3 \times 10^6$ Dalton (<em>E. coli</em>)</td>
<td>$1.9 \times 10^{12}$ Dalton (<em>Human cell</em>)</td>
</tr>
</tbody>
</table>
gives us an estimate of the allowable $\Delta t$ values for different $E_{\text{act}}$'s such that $T_1$ indeed remains constant. With increasing $\Delta t$, the time taken for successive collisions between the TF and DNA increases, resulting in an overall increase in the average binding time. However, with $\Delta t \leq 10^{-8}$, $T_1$ remains constant for each $E_{\text{act}}$.

Fig. 5. $T_1$ against increasing $\Delta t$ for E. coli: Fig. 6. $T_1$ against increasing $E_{\text{act}}$ for E. coli: the average time for binding is generally independent of $\Delta t$.

Fig. 7. $T_1$ against increasing number of binding sites for E. coli: the average time for binding decreases as the number of binding sites increase.

Fig. 8. $T_1$ comparison with experimental results: we report the minimum, maximum and average times for the random DNA-protein binding time where the average time lies in the $1 \sim 10$ secs range.

Fig 6 plots $T_1$ against the different possible $E_{\text{act}}$ estimates. It shows that the average time for binding increases with increasing $E_{\text{act}}$ values. As $E_{\text{act}}$ increases, more kinetic energy is required by the TFs to achieve stable binding, and only higher molecular velocities can produce that energy. Hence $p_b$ decreases resulting in an overall increase in $T_1$. However, for very low $E_{\text{act}}$, the binding times tend to increase because the TFs actually has to spend more time to
bind to a DNA site due to low kinetic energy requirement. Another interesting feature is that $T_1$ remains the same for different estimates of $\Delta t$ as long as $0 \leq p_b \leq 1$. As discussed before, the regions to the left of the vertical lines denote the forbidden regions where $p_b > 1$. The speed-stability paradox [24] says that for acceptable average time estimates we should have $\sigma \sim k_B T$, whereas for stable binding we need $\sigma \geq 5k_B T$. Our results show that we can achieve stable binding between $E_{act} = 1k_B T$ for $\Delta t = 10^{-8}$ s and $E_{act} = 13k_B T$ for $\Delta t = 10^{-4}$ s. The minimum possible values for $E_{act}$ for different $\Delta t$’s are reported in Table 2. The average time for TF-DNA binding is experimentally measured [24] to be $1 \sim 10$ s, which is achieved with $E_{act} \simeq 20k_B T$.

Fig 8 gives the comparison between the experimental results and our theoretical estimates. We find that for $20k_B T \leq E_{act} \leq 26k_B T$, our results match with the experimental values. The minimum and maximum times for binding reported in the figure for different $E_{act}$ values are calculated assuming 95% confidence interval. Thus, our theoretical model also gives an estimate of the activation energy required for stable binding. It should be noted that $E_{act}$ refers to the total free energy change due to binding and should be higher than $\sigma$ as calculated in [24]. We also find that in the range $20k_B T \leq E_{act} \leq 26k_B T$, the time of binding follows an exponential distribution (as the calculated mean is very close to the standard deviation).

In Fig 7, we find that $T_1$ decreases as the number of binding sites $G$ is increased which is again logical as the protein molecules now have more options for binding.
Table 2
Allowable $E_{act}$ values against $\Delta t$ such that $0 \leq p_b \leq 1$

<table>
<thead>
<tr>
<th>$\Delta t$ (in secs)</th>
<th>Minimum $E_{act}$ (in $k_B T$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-4}$</td>
<td>13</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>10</td>
</tr>
<tr>
<td>$10^{-6}$</td>
<td>7.6</td>
</tr>
<tr>
<td>$10^{-7}$</td>
<td>5</td>
</tr>
<tr>
<td>$10^{-8}$</td>
<td>1</td>
</tr>
</tbody>
</table>

5.2.2 Results for eukaryotic cells

Fig 9 shows similar trends for eukaryotic cells. The $T_1$ values for eukaryotic cells are higher than those for bacterial cells mainly because the volume of the nucleus is larger than the average volume for prokaryotic cells. Also, $\alpha$ decreases the probability of binding appreciably as the DNA is arranged in nucleosome cylinders, thereby reducing the average surface area for collision and hence reducing $p_b$. Also, the $p_d$ component of $p_t$ results in lesser values of $p_t$ for eukaryotic cells and hence greater values for $T_1$.

Fig 10 shows the dependence of $T_1$ on $\alpha$. With smaller $\alpha$, the value of $C_{eff}$ is smaller, and hence $T_1$ is higher. It can be observed that $\alpha$ does not significantly affect the average time for binding.

Figs 7, 9, 10 were generated with $E_{act} = 15$ $k_B T$. For eukaryotic cells, we consider the average time for binding after the TF has diffused inside the nucleus. Thus, the overall time for DNA-protein binding has to consider the time taken by protein molecules for diffusion. This has been extensively studied and not reported here.

5.2.3 Important observations from the per = 1 results

(1) Our model achieves the experimental estimate of $1 \sim 10$ secs with activation energy in the range: $20k_B T \leq E_{act} \leq 26k_B T$ for prokaryotic cells (obviously the results are generated for the PurR TF in E. coli and we have not tested this range for other TFs as yet). The corresponding range for eukaryotic cells has not been reported here because we need to know the corresponding experimental estimates for human cells.

(2) The stochastic nature of protein-DNA binding time can be approximated by an exponential distribution in this range as the observed values for
mean and standard deviation of the binding time are comparable.

(3) The average time for DNA-protein binding increases for higher $E_{act}$.

(4) The DNA-protein binding time is independent of the value of $\Delta t$. The recommended value of $\Delta t$ is $10^{-8}$ secs. Figs 5-6 show the dependence of the average time on $\Delta t$ and $E_{act}$. We find that a wider range of $E_{act}$ is available (keeping $p_b \leq 1$) with lesser $\Delta t$. The same estimate holds true for eukaryotic cells also.

(5) The average time decreases as the number of DNA binding sites increase because the TF has more sites to bind to.

(6) The average time is not significantly affected by $\alpha$, i.e., the percentage availability of average collision cross-sectional area.

Fig. 11. CDF of the random DNA-protein binding time for $E_{act} = 22k_BT$, $\Delta t = 10^{-8}$ in *E. coli*.

Fig. 12. $T_1$ measurements with increasing $\Delta t$ for eukaryotic cells.

Fig. 13. Average Time against increasing number of binding sites for eukaryotes.

Fig 11 plots the cumulative distribution function (CDF) for the time of binding with $E_{act} = 22k_BT$ for *E. coli*. Figs 12 and 13 respectively show the dependence of $T_1$ on $\Delta t$ and the number of binding sites for eukaryotic cells.
5.3 Validation of DNA replication with no-sliding assumption

We used another model validation exercise having robust measurement data. We build the DNA replication model of *E. coli* that provides the gross measurement data of a large number of DNA nucleotide/protein interaction sequences. We also build the analytical model from the micro-scale DNA nucleotide/protein interaction times to copy the DNA.

Table 3 presents the parameters used to compute the total time taken for DNA replication using the TF-DNA binding model as the base model and assume that the TFs never slide on the DNA (*per* = 1). We assumed that (i) the rate of replication is the same in both leading strand and lagging strands and (ii) replication stops at the position directly opposite to the OriC in the chromosome. We now estimate the individual time delays in each step of DNA replication mechanism.

1) Binding of DnaA, initiation proteins, with DNA at OriC: We consider length of the replication as 245 bps [28] and 20 molecules of DnaA proteins bind with DNA [28] one after another at OriC. The total time delay for the whole process will be 20 times the time taken to bind one molecule of DnaA with DNA at OriC. With $E_{act} = 20k_B T$, $T = 273K$, $L = 245bps$ and $1bp = 0.34 \times 10^{-9}m$, the time taken for a DnaA is 0.133 secs. Hence, the time taken for 20 molecules of DnaA is $S_1 = 2.6565$ secs.

2) Binding of DnaB (Helicase) with DNA double helix at OriC: DnaB binds with the complex formed by DnaA molecules. Two molecules of DnaB enzymes will be required for one of the two replication forks. We ignored the role of DnaC (another enzyme that helps loading the DnaB with the complex), since the loading function is not known clearly. The time taken for a DnaB is computed as 0.18 secs. Hence, the time taken for 2 molecules of DnaB is $S_2 = 0.36$ secs.

3) Binding of DnaG (Primase) with initiation complex: A molecule of DnaG binds with the complex formed after the previous step. For the two replication forks, two DnaG enzymes will be used. Hence the total time delay is twice the
Table 4
$E_{act}$ and $per$ requirements for $n = 100 bps$

<table>
<thead>
<tr>
<th>$\sigma$ (in $k_B T$)</th>
<th>$E_{act}$ (in $k_B T$)</th>
<th>$per$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>20 – 26</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>20 – 26</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>11 – 15 or 20 – 26</td>
<td>0.1 – 0.9 or 1.0</td>
</tr>
<tr>
<td>2</td>
<td>14 – 17 or 20 – 26</td>
<td>0.1 – 0.9 or 1.0</td>
</tr>
<tr>
<td>1</td>
<td>20 – 24 or 20 – 26</td>
<td>0.1 – 0.9 or 1.0</td>
</tr>
</tbody>
</table>

The time taken to bind one molecule of DnaG with the complex. We compute the time taken for a DnaG as 0.15 secs. Hence, the time taken for 2 molecules of DnaG is $S_3 = 0.3$ secs.

4) Binding of DNA polymerase III holoenzyme (Polymerase) complex with replication formed after step 3 in the DNA double helix: 2 DNA polymerase III holoenzymes are required for the two replication forks. Hence the time delay for this step is twice the time taken for binding one molecule of DNA Polymerase III holoenzyme. We compute the time taken for a DNA Poly Holoenzyme with DNA = 0.363 secs. Hence, the time taken for 2 molecules of DNA Poly Holoenzymes with DNA is $S_4 = 0.726$ secs.

5) Unwinding of DNA by Helicase by hydrolyzing 1 ATP molecule: Helicase unwinds the double stranded DNA by hydrolyzing ATP and the rate of unwinding is 3 bps by hydrolyzing one ATP molecule to ADP. We compute the time taken for unwinding 3 nucleotides = 0.002736 secs. And, time taken to unwind 33 nucleotides, $S_5 = 0.7$ secs.

6) Coating of ssDNA with SSB protein for stabilizing replication process: We assume one SSB molecule covers ~ 33 nucleotides in the ssDNA [28]. SSB proteins are required in both leading as well as lagging strands. These proteins are continuously attached with ssDNAs before the new DNA strand is synthesized and attached. We compute the time taken for coating 3 nucleotides as 0.002736 secs. Hence, the time taken to coat 33 nucleotides is $S_6 = 0.7$ secs.

7) Synthesis of new DNA by DNA polymerase III holoenzyme: DNA polymerase III synthesize new DNA at the rate of 3 nucleotides [28] by hydrolyzing 1 ATP molecule to ADP. The time taken to synthesize 3 nucleotides by DNA Poly III holoenzyme is computed as 0.00275 secs. Therefore, the total time required for the complete DNA is 35.403 min.

Adding the time delays from each of the above steps, the total time required for DNA replication in E. Coli from our model is ~ 36 mins which is quite close to the experimental estimate of 42 mins.
5.4 Numerical results for the combined model in E. coli when \( \text{per} \neq 1 \)

In [24], the authors presented an experimental estimate of \( \tau_{id} \) for different values of sliding distance (denoted by \( \pi \)) and at different roughness \( \sigma \) for the PurR TF of \textit{E. Coli} with a random and uncorrelated energy profile having standard deviation \( \simeq 6.5k_BT \). These \( \tau_{id} \) estimates have been used to generate the plots.

![Fig. 14. Average binding time for purR](image1)

\( (\sigma = 1k_BT) \)

![Fig. 15. Average binding time for purR](image2)

\( (\sigma = 2k_BT) \)

![Fig. 16. Average binding time for purR](image3)

\( (\sigma = 3k_BT) \)

![Fig. 17. Average binding time for purR](image4)

\( (\sigma = 4k_BT) \)

Figs 14-18 plot \( T_1 \) for \( \sigma = 1, 2, 3, 4, 5 \) \( k_BT \) respectively with \( \text{per} = 0 \) and different values of the sliding distance, \( n \), in bps. The x-axis gives the values for \( E_{act} \) and the y-axis is plotted on a logarithmic scale with \( E \pm z = 10^{\pm z} \). Note that the average binding time estimates increase with increasing \( \sigma \).

For \( \sigma = 1k_BT \) and \( \text{per} = 0 \), the experimental estimates of 1 ~ 10 secs can be achieved with \( 15k_BT \leq E_{act} \leq 20k_BT \), even with \( n = 8000 \text{bps} \). However, the experimental results can be achieved up to \( (n = 2000 \text{bps}, \sigma = 2k_BT), (n = 200 \text{bps}, \sigma = 3k_BT), (n = 20 \text{bps}, \sigma = 4k_BT) \) and \( (n = 7 \text{bps}, \sigma = 5k_BT) \).
Fig. 18. Average binding time ($\sigma = 5k_B T$).

Table 5
$E_{act}$ and $\text{per}$ requirements for $n = 50$bps

<table>
<thead>
<tr>
<th>$\sigma$ (in $k_B T$)</th>
<th>$E_{act}$ (in $k_B T$)</th>
<th>$\text{per}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>20 − 26</td>
<td>1.0</td>
</tr>
<tr>
<td>4</td>
<td>20 − 26</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>12 − 15 or 20 − 26</td>
<td>0.1 − 0.9 or 1.0</td>
</tr>
<tr>
<td>2</td>
<td>20 − 24 or 20 − 26</td>
<td>0.1 − 0.9 or 1.0</td>
</tr>
<tr>
<td>1</td>
<td>22 − 25 or 20 − 26</td>
<td>0.1 − 0.9 or 1.0</td>
</tr>
</tbody>
</table>

Thus if we assume that every collision of the TF with the DNA is accompanied with a 1-d diffusion, the average number of base pairs that the TF can slide is only 7 bps when $\sigma = 5k_B T$. This is certainly a very low estimate and it is logical to assume that not every TF-DNA collision involves 1-d diffusion.

The next step is to find an estimate of $\text{per}$ ($\neq 0$), that gives binding times in the experimental range even with biologically relevant amounts of sliding. In [24], the authors reported the optimal number of base-pairs that can be searched at $\sigma = 1k_B T$ as 100 bps. We report the maximum $\sigma$ that can achieve the experimental estimates from our results in Table 4 and that for 50 bps in Table 5. Thus we can get the bounds on $E_{act}$, for different combinations of $\text{per}, \sigma$ and $n$. The above results show the maximum value of $\sigma$ for which the experimental rate can be achieved. However, for $\sigma = 5k_B T$, we have to consider either $\text{per} = 1.0$, i.e., the TF does not slide on the DNA, or it can slide a maximum of 7 bps.
5.5 Simulating the dynamics of protein-DNA binding

In this section, we analyze the dynamics of the protein-DNA binding event at a “systems level” - studying its effect in association with other molecular events involved in a cellular process. In particular, we focus on the effects of TF binding event on the expression of genes in prokaryotic cells.

Stochasticity in prokaryotic gene expression has been extensively studied, both mathematically [30],[31], as well as in experimental systems [32],[33]. Particularly, the *burstiness* in protein production i.e., proteins are produced in random bursts of short duration rather than in a continuous manner, have been shown in single cell experiments conducted on the *lacZ* gene in *E. Coli* [32],[33]. The random fluctuations in the number of proteins, termed ‘noise’, stems from the interplay of a large number of factors: discrete, random nature of molecular interactions like RNA Polymerase (RNAP) - promoter binding and transcription open-complex formation, low copy number of key transcriptional and translational machineries like RNA polymerase, transcription factors, ribosomal units etc. and the random nature of signals triggering gene expression. The fine-grained regulation of gene expression by the transcriptional machinery, specifically, transcription initiation frequency controlled by the binding of the transcription factor (TF) upstream of the promoter region, has been quantitatively studied in [34],[35],[36].

In order to quantitatively study the stochastic dynamics of TF-DNA binding on prokaryotic gene expression, we build a discrete-event based simulation environment, as outlined in [7], capturing the key molecular events involved in the process.

- **Transcription Event:** This event represents the triggering of transcription by the activation of a gene and the eventual release of a mRNA molecule in the system. The probability distribution characterizing the time taken for the event is defined by its first and second moments, $\bar{R}_{\text{mR}}$ and $\sigma_{\text{mR}}$ respectively, and the time between two transcription events is represented by the random variable $\tau_{\text{transcription}}$. This event encompasses the micro-event of TF-DNA binding and includes the average binding time mathematically captured in $T_1$ and $T_2$.

- **Transcript Decay Event:** This event represents the decay of a transcript and is characterized by an exponential distribution with half-life $m_{\text{decay}}$ obtained from experimental data [37].

- **Translation Event:** This event captures the process of protein synthesis from a single mRNA molecule characterized by the probability distribution of its time ($\bar{R}_{p}$ and $\sigma_{p}$)

---

2 The details of the stochastic models for prokaryotic transcription and translation, together with the simulation framework are available in [37]
Protein Decay Event: This event represents the decay of a protein characterized by an exponential distribution with half life of $m_{protein}$ obtained from experimental data [37].

The interactions of these molecular events, as captured in Fig 19, drives the dynamics of protein production in prokaryotic cells. In order to study the effect of TF-DNA binding time, as expressed by the parameterized model elucidated in the previous section, on the stochasticity of protein synthesis, we conducted several in silico experiments by varying the average binding time for the TF-DNA binding microevent involved in transcription $^3$.

5.5.1 Protein synthesis dynamics with TF-DNA binding time of 10 secs

We conducted simulation studies to validate experimentally observed “bursts” in protein generation of E.Coli. With the TF-DNA binding time of 10 secs (based on experimental observations reported in the previous section), Fig 20 shows the temporal dynamics of mRNA and protein molecules together with the noise profile (noise being quantitatively measured as the ratio of the variance to squared mean [30],[37]. As observed from the plots, the burstiness in the number of lacZ proteins produced (marked by a corresponding increase in $^3$

\textsuperscript{3} The simulation was carried for the lacZ gene expression in E.Coli to validate with available experimental data. The simulation experiments were conducted for 10 cell cycle times and results represent average value for 50 simulation runs.
noise) is primarily caused by the low frequency of transcription events (around 1.2 mRNA molecules are produced per cell cycle).

Fig. 20. Dynamics of lacZ gene expression with experimental TF-DNA binding time

5.5.2 Protein synthesis dynamics with TF-DNA binding time of 0.1 secs

As noted in [34], the transcription initiation frequency has a key role in controlling the nature of stochasticity in protein synthesis. In order to analyze the impact of the TF-DNA binding event in this fine-grained regulation, we conducted simulation experiments for the lacZ system with different average TF-binding times computed from our model. In Fig 21, we show the dynamics of the gene expression process for TF-DNA binding time of 0.1 secs. As seen from the figure, a decrease in the average binding time does not significantly increase the transcription event rate as observed in the similar protein and noise profiles as reported in the previous simulation case study.
5.5.3 Protein synthesis dynamics with TF-DNA binding time of 100 secs

In order to further analyze the effect of the TF-DNA binding event, particularly with increased event time, we conducted a simulation experiment with the TF-DNA binding event time set to 100 secs, an order of magnitude higher than the experimentally reported value. As seen from the protein and noise profiles in Fig 22, the TF-DNA binding time causes the transcription event time to decrease (i.e., the number of mRNA molecules released decreases due to the high TF-DNA binding time), thus increasing the “burstiness” in subsequent protein synthesis.

In this section, we have quantitatively captured the effect of the TF-DNA binding event as part of a dynamical system involving the temporal interaction of multiple molecular events associated with gene expression in prokaryotic cells. Our simulation results confirm biological observed burstiness in protein synthesis while providing in silico insights into the role of TF-DNA binding on the amplitude of fluctuations (noise) of the gene expression process.
5.6 Limitations of our model

Maxwell-Boltzmann distribution of molecular velocities: The Maxwell-Boltzmann distribution gives a good estimate of atomic/molecular velocities and is widely used in practice. Molecular dynamic (MD) simulation measurements during protein reactions show that the velocity distribution of proteins in the cytoplasm closely match the Maxwell-Boltzmann distribution. However, its application in our collision theory model might not give perfect results. Ideally the velocity distribution should incorporate the properties of the cytoplasm (nucleus for eukaryotes). We plan to extend our model to incorporate more realistic velocity distributions in the future.

3-d protein structure: The $p_m$ estimation can be improved by considering the 3-d structure of the protein. Ideally, the motif of the protein molecule is located towards the outer surface such that $p_m$ is actually higher than what we compute.

The actual protein-DNA binding process: The present model does not
incorporate the time required for the actual binding process i.e., how the specific atoms of the protein form chemical bonds with the DNA forming a stable complex. Also, it should be noted that the complex machinery of transcription, especially for eukaryotes, is not completely understood, yet. Many proteins can play a role in regulating one gene that would require further analysis. Our model can serve as a starting point for handling such cases.

**Nucleosome dynamics:** The long DNA chain in eukaryotes uses a systematic hierarchical compression. In the lowest compaction level the genetic material comprises arrays of coiled DNA around histones (globular octamer of cationic nucleus proteins) [28]. Each of these array elements is called a nucleosome that exhibits the following four dynamics: (1) compositional alternation, (2) covalent modification, (3) translational repositioning, and (4) conformational fluctuation. Compositional alternation is done by some remodeling enzymes to promote gene activation. Post translational modifications including acetylation, methylation, phosphorylation and ubiquitination are among the covalent modifications that can destabilize the histone cores and exploit DNA access to the biological processes. ATP-dependent remodellers use energy derived from ATP hydrolysis to loosen the contacts between the coiled DNA and the histone core. In translational repositioning, the bp position of core particles in the genome change to enhance the target site access. This process can happen both intrinsically or by the aid of remodellers. Conformational fluctuation is a periodic minor change to the conformation of a canonical nucleosome. The model presented in this chapter can help incorporate these factors in a more comprehensive protein-DNA binding model.

### 5.7 Biological implications

**Several TFs searching simultaneously:** If we consider several TFs searching for their sites on the DNA simultaneously, our results still remain valid. In [24], the authors argue that this may reduce the total search time because the experimental estimate of $1 \sim 10$ secs is generated from the binding rate of the TFs to the DNA site. Our results, however, are for any specific TF and compute the average time required for this TF to bind to the DNA. Thus increasing the number of TFs should not change the results that we report for any particular TF. In fact, this brings down the experimental estimate of the binding time and hence requires lesser $E_{act}$ for stable binding (as discussed in the next subsection). This may also cause molecular crowding in the cell which can have an impact on the search time. We did not consider molecular crowding on DNA or protein hopping (intersegment transfer) in our model for similar reasons as in [24].

**Funnels and local organization of sites:** Several known bacterial and eukaryotic sites tend to cluster together. Such clustering or other local arrangement of the sites can create a funnel in the binding energy landscape
leading to more rapid binding of cognate sites. Our model assumes no such
funnels of energy field. In the present model, the probability of collision is
assumed uniform for the entire DNA. Because of local organization of sites,
there is bias in the collision site, we can model that effect by changing the
uniform distribution by another distribution to represent this bias. Also, due
to change in energy landscape if the binding energy requirement changes, the
probability of binding will increase in our model and hence will reduce the
binding time.

Possible experiments to test our predictions: The search time depends
on the activation energy of the TFs, which, in turn, can be controlled by the
ionic strength of solution. Also, we show how the binding rate depends on the
average collision time between two random segments of DNA, \( \tau \). This time
measurement \( (\tau) \) depends on the DNA concentration and the domain orga-
nization of DNA. By changing DNA concentration and/or DNA stretching
in a single molecule experiment, one can alter \( \tau \) and thus study the role of
DNA packing on the rate of binding. This effect has implications for DNA
recognition in vivo, where DNA is organized into domains. Similarly, one can
experimentally measure and compare the binding rate, in the presence of other
DNA-binding proteins or nucleosomes.

Biological relevance of our model: Our model suggests that the kinetic en-
ergy of the TFs has to exceed \( E_{act} \) for successful binding. Is the kinetic energy
of the TFs greater than this minimum requirement in general? Theoretically,
of course, the energy can be infinitely large for any molecule. Moreover, the
bound on \( E_{act} \) can be brought down significantly if we incorporate the above
factors. Note that the experimental estimate of 1 ~ 10 secs incorporates the
actual binding time. Thus the time for searching a DNA site by a TF should
be quite smaller than 1 ~ 10 secs resulting in a very low requirement of \( E_{act} \).
Also, because the experimental results depend on the binding rate, the total
search time for 100 copies of a TF searching in parallel for the cognate site
in a cell of 1 \( \mu m^3 \) volume is \( \approx \) 0.1 s. This estimate further decreases with
increasing number of TFs. So, to compute the average time for binding experi-
mentally, we really need to compute the average number of that particular TF
in the cell. Thus the model presented in this chapter can be further extended
to incorporate these factors.

We have implemented a discrete event simulation framework for the PhoPQ
signal transduction system in Salmonella Typhimurium [7]. The framework
uses the above protein-DNA binding time estimates (with per=0, i.e. the no
sliding assumption) combined with the execution time estimates of some other
elementary biological events e.g. cytoplasmic reactions [8,11], diffusion based
molecular transport [9] and ligand-protein docking [10]. The corresponding
system simulation results can be found at:
http://crewman.uta.edu/dynamic/bone/projects.htm
6 Conclusion

We have presented a simplified model to estimate the DNA-protein binding time by transforming the biological function as a stochastic process of a number of biological microevents. The probabilities of these microevents are used to create the complete stochastic model of the biological event. We used collision theory and Maxwell Boltzmann velocity distribution to calculate this microevent probability. The model is computationally fast and provides two moments for this random binding time. The model is robust as the major factors are captured for general cell environments. The complexity of DNA packing has been simplified to achieve acceptable estimates of the DNA-protein binding time. We found the range of activation energies of the TFs that are crucial for the robust functioning of gene transcription. The speed-stability paradox can also be bypassed using the no TF sliding assumption and its effects reduced if we incorporate 1-d diffusion. The proposed mechanism has important biological implications in explaining how a TF can find its site on DNA in vivo, in the presence of other TFs and nucleosomes and by a simultaneous search by several TFs. Beside providing a quantitative framework for analysis of the kinetics of TF binding (and hence, gene expression), our model also links molecular properties of TFs and the location of the binding sites on nucleosome cylinders to the timing of transcription activation. This provides us with a general, predictive, parametric model for this biological function. Thus, our discrete stochastic modeling can incorporate more parameters in the simulation.

7 Acknowledgements

The authors thank L. Lekhendro Singh (University of Texas at Arlington, Email: lsingh@cse.uta.edu) for validating the model with DNA replication data in *E. Coli*. We also thank the anonymous reviewers for their suggestions that significantly improved the quality of this paper.

References


and chemical significance revealed by single-molecule measurements. J. Biol.

of protein translocation on nucleic acids. 1. Models and theory. Biochemistry,

regulatory proteins. Statistical-mechanical theory and application to operators

of protein translocation on nucleic acids. 3. The Escherichia coli Lac repressor-
operator interaction: kinetic measurements and conclusions. Biochemistry, 1989,
20:6961-6977.


Target Location in Sequence-Dependent Potential. Biophysical Journal, 2004,
87:4021-4035.

Paulsen, M. Peralta-Gil and P. D. Karp. EcoCyc: A comprehensive database

N. Shindyalov, P. E. Bourne. The Protein Data Bank. Nucleic Acids Research,


0815316194.

[29] G. D. Stormo and D. Fields. Specificity, free energy and information content in

2:157-75.

[31] H. McAdams and A. Arkins. Stochastic mechanisms in gene expression. PNAS,

[32] L Cai. Stochastic protein expression in individual cells at the single molecule


