Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Modeling protein-DNA binding time in Stochastic Discrete Event Simulation of Biological Processes

Preetam Ghosh, Samik Ghosh, Kalyan Basu, and Sajal Das
Biological Networks Research Group (BONE), The University of Texas at Arlington, USA.
Email: {ghosh, sghosh, basu, das}@cse.uta.edu

***Abstract*— This paper presents a parametric model to estimate the DNA-protein binding time using the DNA and protein structures and details of the binding site. To understand the stochastic behavior of biological systems, we propose an "in silico" stochastic event based simulation that determines the temporal dynamics of different molecules. This paper presents a parametric model to determine the *execution time* of one biological function (i.e. simulation event): protein-DNA binding by abstracting the function as a stochastic process of microlevel biological events using probability measure. This probability is coarse grained to estimate the stochastic behavior of the biological function. Our model considers the structural configurations of the DNA, proteins and the actual binding mechanism. We use a collision theory based approach to transform the thermal and concentration gradients of this biological process into the probability measure of DNA-protein binding event. 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 to be used by 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.**

I. 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 [6] specially for protein creation and other signaling pathways. The stochastic simulation models [7], [8], [9], [10], [11] 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, and hence not suitable for complex signaling pathway problems. Our goal is to build a stochastic discrete event based framework [5] for biological systems to overcome the computational complexity of current mesoscale and stochastic simulation methods. This flexible simulation framework can also 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 event models are used to develop a stochastic discrete-event simulation. The event modeling uses an 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 [5] and the abstraction mechanisms in [13], [14], [15], [16]. Here, we extend the event modeling approach to compute the execution time of another complex biological function: 'DNA-protein' binding.

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 from experiments. The classical protein-DNA sliding model considers the energetics of protein-DNA interactions [4]. In contrast to the existing thermodynamic and diffusion based models, our approach closely follows the biological process divided into discrete microevents. 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 for our stochastic simulator (note that the Gillespie simulator uses a simple rate constant 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.

II. DNA-Protein Binding Model

We partition this problem into 2 biological microevents: 1) 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 2) 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.

1-4244-0710-9/07/$20.00 ©2007 IEEE

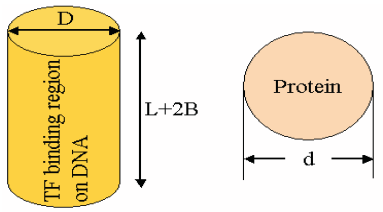

Fig. 1. Schematic diagram of protein molecule and TF binding region of the DNA

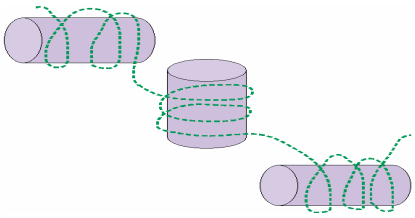

Fig. 2. DNA packing through nucleosomes

A. *Modeling the first microevent: Calculating p_n*

In this section we abstract the first microevent by computing the probability of collision of the protein (TF) with the binding site ($\pm B$) on the DNA (denoted by p_n). 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.

We define our coordinate system such that the DNA is stationary with respect to the protein molecule, such that the latter moves towards the DNA with a 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 Δt 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.

Calculating the average surface area of collision of 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 cylinder, and (3) cylinder at an arbitrary angle, θ, with the direction of motion of the protein (Fig 3). The cross-sectional area of collision, C, is given by:

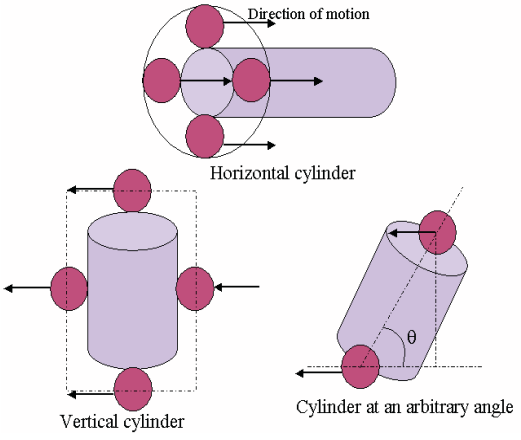

Fig. 3. Collision of spherical protein and cylindrical DNA transcription factor binding region.

$$C = \left\{ \begin{array}{ll} \pi\frac{(d+D)^2}{4}, & \text{for } \theta = 0^0 \\ (L+2B+d)(D+d), & \text{for } \theta = 90^0 \\ (D+d)(L+2B+d)\sin\theta, & \text{otherwise} \end{array} \right\}$$

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 the whole calculation can be limited to the case where $0^0 < \theta < 90^0$. We assume an 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:

$$C_{avg} = \int_0^{\frac{\pi}{2}} \frac{2}{\pi} C(\theta) d\theta = \frac{2}{\pi}(D+d)(L+2B+d).$$

This cross-section C_{avg}, moves in the cytoplasmic space (nucleus for eukaryotes) to create the collision volume for a particular binding site.

Probability of protein-DNA binding in eukaryotic cells: Fig 2 illustrates how DNA is packed along different cylindrical nucleosomes. Thus, L in C_{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 [12] 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.

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_h^c. We have:

$$\begin{aligned} p_h^c &= \frac{\text{length of that nucleosome cylinder}}{\text{length of all nucleosomes + length of all stretches}} \\ &= \frac{l_n}{N_1 l_n + \sum_{i=1}^{N_2} l_s^i} \end{aligned}$$

where, l_n denotes the length of a nucleosome cylinder (assumed fixed for all the cylinders), l_s^i 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. N_1 and N_2 are the number of nucleosome cylinders and that of stretches of DNA respectively. Now, the probability of hitting the DNA portion of the nucleosome cylinder, p_d can be estimated from the surface area of the nucleosome cylinder and that of the DNA present in the cylinder as follows:

$$p_d = \frac{\pi D l_d}{\pi D l_d + \pi d_n l_n}$$

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 $\frac{l_d}{l_n} = 1.65$. And, p_f^c designates the probability of colliding with the TF binding region ($\pm B$) 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 of the TF binding region ($\pm B$) of the DNA for a binding to occur. So, we need to calculate the probability of identifying the motif of the colliding protein molecule, p_m 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 ($\pm B$) 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. We approximate this case through a difficulty parameter α, which denotes the *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: $C_{eff} = \alpha \times C_{avg}$

Case II: In this case, the probability of hitting the correct stretch of DNA in between the nucleosome cylinders is designated by p_h^s as follows:

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

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:

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

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: $C_{eff} = C_{avg}$

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^c 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^c and p_f^s computations should change as follows:

$$p_f^c = \frac{\text{length of TF binding region portion in nucleosome + B}}{\text{total length of DNA in that particular nucleosome}}$$

$$p_f^s = \frac{\text{length of TF binding region portion in the stretch + B}}{\text{total length of DNA in that particular stretch}}$$

And hence we have:

$$p_w^c = p_m \times p_h^c \times p_f^c \times p_d; \quad p_w^s = p_m \times p_h^s \times p_f^s; \quad p_n = p_w^c + p_w^s$$

Thus total probability of collision of the TF to the DNA binding site ($\pm B$) is p_n. Also, the average cross-sectional area calculations become a little different in this case. We break up C_{avg} into C_{avg_1} and C_{avg_2} based on L_1 and L_2, where, L_1 is the length of the TF binding region in the nucleosome cylinder and L_2 denotes that in the adjoining stretch. 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 DNA packed inside a nucleosome cylinder. However, if the region 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: $C_{eff} = \alpha \times C_{avg_1} + C_{avg_2}$

Thus the total probability of collision to one specific TF binding region, p_n, 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 in nucleosome cylinders is fixed and hence we can find where the TF binding region is located as described in Cases I, II or III.

Approximate mechanism of finding where the TF binding region is located: 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 (πd_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. Also, we assume that all the stretches of DNA between nucleosome cylinders are equal in length. Thus, we can approximate the length of DNA in a stretch 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 in the above expression is $(N-1)$ because we assume that there can only be $(N-1)$ stretches of DNA present in between the N nucleosome cylinders. Also 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.

Protein-DNA binding probability for bacterial cells: The bacterial genome is supercoiled with a general organization as depicted in Fig. 4. Each domain consists of a loop of DNA, the ends of which are secured in some way. Hence, the total

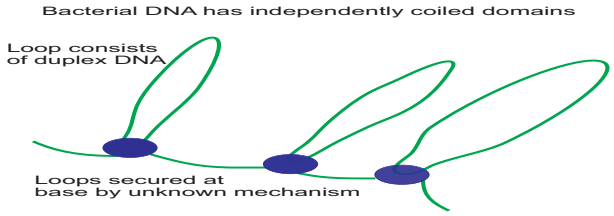

Fig. 4. Bacterial Genome Structure.

probability of collision in this case is simply approximated as:

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

Also, because the entire surface area of the DNA is available for binding, the effective cross-sectional area of binding is given by: $C_{eff} = C_{avg}$

B. 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. Thus, p_b is the information domain abstraction (in terms of probability) of the second microevent. In time Δt, the TF sweeps out a volume ΔV given by:

$$\Delta V = C_{eff} U \Delta t$$

Let the total volume of the cell be V (for a prokaryotic cell, we 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 as we assumed the movement of the TF is confined within the nucleus at this stage). We next assume that the colliding protein molecule must have free energy E_{Act} or greater to bind to the specific DNA TF 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 relative velocity U is $E = \frac{m_{PD}U^2}{2}$, where $m_{PD} = \frac{m_P . m_D}{m_P + m_D}$ = the reduced mass, m_P = mass (in gm) of the protein molecule and m_D = mass (in gm) of the DNA. U reflects the cumulative effects of all the force fields on the mass of the protein and we approximate this complex dynamic process by a statistical distribution to capture the uncertainty represented by the Maxwell-Boltzmann distribution of molecular velocities for a species of mass m given by:

$$f(U,T)dU = 4\pi(\frac{m}{2\pi k_B T})^{3/2} e^{\frac{-mU^2}{2k_B T}} U^2 dU$$

where k_B = Boltzmann's constant = 1.381×10^{-23} $kg\,m^2/s^2/K/molecule$ and T is the absolute temperature (= 273 K). We also assume that as the kinetic energy, E, increases above E_{Act}, the number of collisions that result in binding also increases. Thus following the concept shown in [13] we get:

$$p_b = \frac{C_{eff}\Delta t}{V}\sqrt{\frac{8k_B T}{\pi m_{PD}}}\, e^{\frac{-E_{Act}}{k_B T}}$$

C. The total binding probability considering all different TF binding regions of the specific protein molecule

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 that is colliding with the DNA. Also, let p_t^i denote the total probability of binding (combining the first and second microevents) for the i^{th} TF binding region ($1 \le i \le G$). 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_n^i and p_b^i 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 $p_n^i = p_n^j$ and $p_b^i = p_b^j$, $i \neq j, 1 \le i, j \le G$. Hence,

$$p_t^i = p_n^i \times p_b^i$$

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

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

This is because, the probability of binding to the first TF binding region is given by $p_t^1 \prod_{j=2}^{G}(1-p_t^j)$; that for the second region is $[p_t^2(1-p_t^1)(1-p_t^3)(1-p_t^4)...(1-p_t^G)]$; and so on. The total probability, p, is the sum of all these individual cases. Thus, p gives us the information domain measure of the complete DNA-protein binding event in terms of probability for any specific TF.

III. Time taken for Protein-DNA binding

We next estimate the time taken to complete the binding with total binding probability p. Let $\Delta t = \tau =$ 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 τ 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 ($\le B$ 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 [4] (where they assume that the TF slides along the DNA at every round). In [4], the authors derived the 1-d diffusion time, τ_{1d} (along the DNA) using the mean first passage time (MFPT) from site 0 to B as follows:

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

where, ν is the effective attempt frequency for hopping to a neighboring site and σ is the roughness of the DNA landscape in units of $k_B T$. Note that τ_{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*. Thus the total probability of binding is:

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

where, $p_{no-sliding}$ denotes the probability of binding when the sliding along the DNA is not considered altogether. Hence, the average time for protein-DNA binding model is given by:

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

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

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

where T_2 is the second moment of the binding time. We find that the time for DNA-protein binding *when no sliding is considered*, follows an exponential distribution for most ranges of E_{act} (reported in the next section). It should be noted that as we assume τ 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.

IV. Results and Analysis

Problems in validation of our model: Before presenting the results, we first discuss the difficulty of experimentally validating our model. Note that we compute the average time for protein-DNA binding in this paper. Existing experimental results are based on estimation of the binding rate of any specific TF to the DNA. And the experimental estimate of $1 \sim 10$ secs is reported from this rate measurement [4]. Hence, the number of TFs in the cell will affect this estimate of time taken by one single TF to bind to the DNA site. 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 next assume that the time taken for any particular 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.

Numerical Results for $per = 1$ (i.e. no TF sliding is considered): In this section, we present the numerical results for the theoretical models derived in the paper. Figs 5-8 present the results for the PurR TF (having 35 binding sites) on the *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 I. We used the EcoCyc database [1] for the E. coli data, and the PDB database [2] for human cell data.

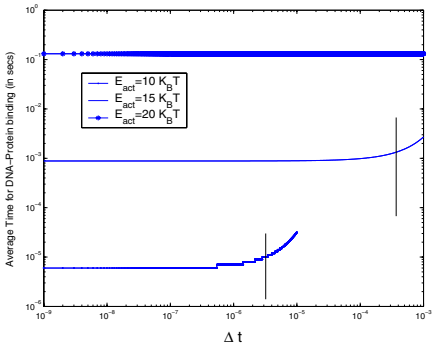

Fig. 5. T_1 against increasing Δt for E. coli.

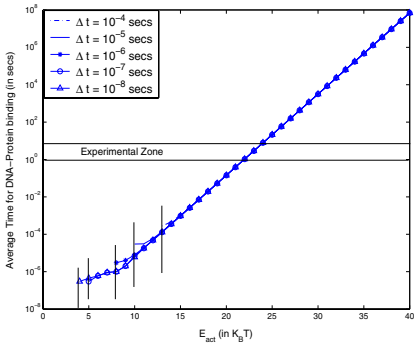

Fig. 6. T_1 against increasing E_{act} for E. coli.

Fig 5 plots T_1 against different values for Δt. The average time for DNA-protein binding remains constant initially and shoots up exponentially with increasing Δt. The same characteristics are seen for different activation energies, $E_{act} = 10$ k_BT, 15 k_BT and 20 k_BT. 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 [3]. Lesser the

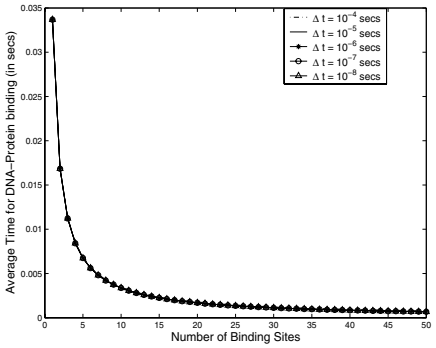

Fig. 7. T_1 against increasing number of binding sites for E. coli.

TABLE I
PARAMETER ESTIMATION FOR BACTERIAL AND EUKARYOTIC CELLS.

Parameters	Prokaryotic Cell	Eukaryotic Cell
V	$4.52 \times 10^{-18} m^3$ (volume of cell)	$4.187 \times 10^{-16} m^3$ (volume of nucleus)
Length of DNA	4.64×10^6 bp (*E. coli*)	3×10^9 bp (*Human cell*)
G	35 (for PurR)	35 (assumed for Htrf1)
Length of TF binding site (L)	26	48
Length of protein amino acid chain	341 (for PurR)	53 (Htrf1)
Length of protein motif	26 (for PurR)	48 (Htrf1)
Radius of Amino acid chain	1 nm (for PurR)	1 nm (Htrf1)
Average radius of the protein ($\frac{d}{2}$)	5 A^o (for PurR)	5 A^o (Htrf1)
m_P	38.175 Dalton (for PurR)	6635 Dalton (for Htrf1)
Diameter of DNA (D)	2 nm (for *E. coli*)	2 nm (*Human cell*)
m_D	3×10^6 Dalton (for *E. coli*)	1.9×10^{12} Dalton (*Human cell*)

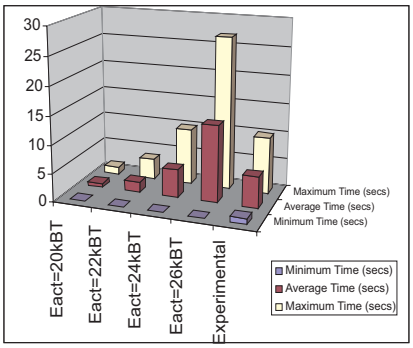

Fig. 8. T_1 comparison with experimental results.

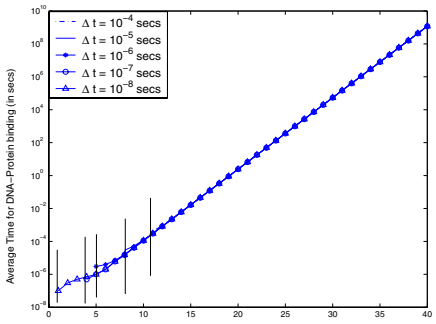

Fig. 9. T_1 against E_{act} for eukaryotes.

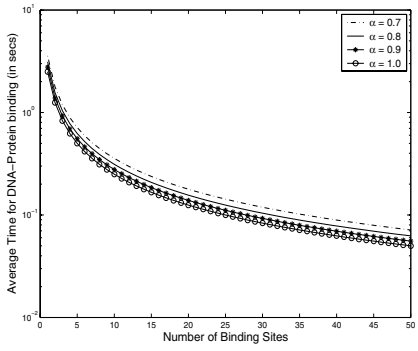

Fig. 10. T_1 against different α's in eukaryotes.

required E_{act}, more is p_b for the protein molecules, and hence lesser is T_1. It is to be noted that p_b as calculated above also corresponds to the number of collisions in time Δt of the protein molecule with the DNA. And for our assumption of at most one collision taking place in Δt 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_{act} plot denotes the forbidden region where $p_b > 1$ even though $0 \leq p \leq 1$. This gives us an estimate of the allowable Δt values for different E_{act}'s such that T_1 indeed remains constant. Note that with increasing Δt, the time taken for successive collisions between the TF and DNA increases, resulting in an overall increase in average binding time. However, with $\Delta t \leq 10^{-8}$, T_1 remains constant for each E_{act}.

Fig 6 plots T_1 against the different possible E_{act} estimates and we find that the average time for binding increases with increasing E_{act} values. As E_{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, with very low E_{act} requirement, we find the binding times tend to increase. This is because the kinetic energy requirement becomes so low, that the TFs actually has to spend more time to bind to a DNA site. Also, an interesting feature is that T_1 remains the same for different estimates of Δt 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 [4] 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 Δt's are reported in Table II. The average time for TF-DNA binding is experimentally measured [4] to be $1 \sim 10s$, 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

TABLE II
ALLOWABLE E_{act} VALUES AGAINST Δt SUCH THAT $0 \leq p_b \leq 1$

Δt (in secs)	Minimum E_{act} (in k_BT)
10^{-4}	13
10^{-5}	10
10^{-6}	7.6
10^{-7}	5
10^{-8}	1

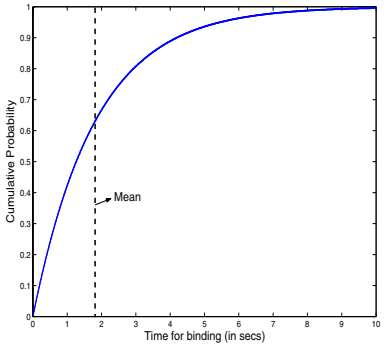

Fig. 11. CDF of our stochastic model for $E_{act} = 22k_BT$, $\Delta t = 10^{-8}$.

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 σ as calculated in [4]. We also find that in the range $20k_BT \leq E_{act} \leq 26k_BT$, 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.

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, α 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 α. With less α, lesser is C_{eff}, and hence higher is T_1. It can be observed that α does not affect the average time for binding significantly.

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 show the dependence of T_1 on Δt and number of binding sites respectively for eukaryotic cells.

Figs 7,9,10 were generated with $E_{act} = 15\ k_BT$. 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.

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: $20k_BT \leq E_{act} \leq 26k_BT$ 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 is approximately independent of Δt and increases for higher E_{act}.
4) An acceptable estimate of Δt is 10^{-8} secs. Figs 5-6 show the dependence of the average time on Δt and E_{act}. We find that a wider range of E_{act} is available (keeping $p_b \leq 1$) with lesser Δt. 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 α i.e. the percentage availability of average collision cross-sectional area.

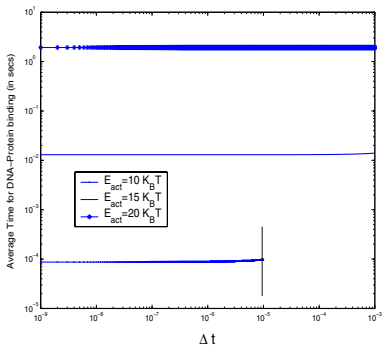

Fig. 12. T_1 measurements with increasing Δt for eukaryotic cells.

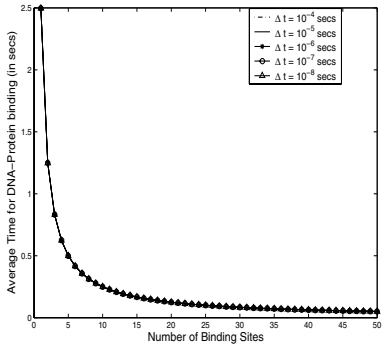

Fig. 13. Average Time against increasing number of binding sites for eukaryotes.

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 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. In E. Coli cells, replication of the single circular chromosome takes about 42 minutes and

TABLE III

E_{act} AND per REQUIREMENTS FOR $n = 100bps$

σ (in k_BT)	E_{act} (in k_BT)	per
5	$20-26$	1.0
4	$20-26$	1.0
3	$11-15$ or $20-26$	$0.1-0.9$ or 1.0
2	$14-17$ or $20-26$	$0.1-0.9$ or 1.0
1	$20-24$ or $20-26$	$0.1-0.9$ or 1.0

TABLE IV

E_{act} AND per REQUIREMENTS FOR $n = 50bps$

σ (in k_BT)	E_{act} (in k_BT)	per
5	$20-26$	1.0
4	$20-26$	1.0
3	$12-15$ or $20-26$	$0.1-0.9$ or 1.0
2	$20-24$ or $20-26$	$0.1-0.9$ or 1.0
1	$22-25$ or $20-26$	$0.1-0.9$ or 1.0

our analytical model predicts the time as $\sim$ 36 mins.

Numerical Results for the combined model in E. coli with $per \neq 1$: In [4], the authors present an experimental estimate of τ_{1d} for different values of sliding distance (denoted by $\overline{n}$) and at different roughness σ for the PurR TF of *E.Coli* with a random and uncorrelated energy profile having standard deviation $\simeq 6.5k_BT$. These τ_{1d} estimates have been used to generate the plots.

For $\sigma = 1k_BT$ and $per = 0$, the experimental estimates of $1 \sim 10$ secs can be achieved with $15k_BT \leq E_{act} \leq 20k_BT$, even with $n = 8000bps$. However, the experimental results can be achieved up to ($n = 2000bps, \sigma = 2k_BT$), ($n = 200bps, \sigma = 3k_BT$), ($n = 20bps, \sigma = 4k_BT$) and ($n = 7bps, \sigma = 5k_BT$). 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 $\sigma = 5k_BT$. 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 per ($\neq 0$), that gives binding times in the experimental range even with biologically relevant amounts of sliding. In [4], the authors report the optimal number of base-pairs that can be searched at $\sigma = 1k_BT$ as 100 bps. We report the maximum σ that can achieve the experimental estimates from our results in Table III and that for 50 bps in Table IV. Thus we can get the bounds on E_{act}, for different combinations of per, σ and n. The above results show the maximum value of σ for which the experimental rate can be achieved. However, for $\sigma = 5k_BT$, we have to consider either $per = 1.0$, i.e. *the TF does not slide on the DNA*, or it can *slide a maximum of 7 bps*.

V. 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 micro events and use the microevents probability information to create the complete stochastic model of the biological event. We used collision theory and Maxwell Boltzmann velocity distribution to get this microevent information. The model is computationally fast and provides two moments for this random number. The model is robust as the major factors are captured in a reasonably accurate way 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. These details make the model more versatile compared to the current rate constants used in the Gillespie simulation. Thus, our discrete stochastic modeling can incorporate more parameters in the simulation.

REFERENCES

[1] EcoCyc: Encyclopedia of Escherichia coli K12 Genes and Metabolism. *http://ecocyc.org/*

[2] The RCSB Protein Data Bank. *http://www.rcsb.org/pdb/*

[3] G. D. Stormo and D. Fields, Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 1998, 23:109-113.

[4] M. Slutsky and L. A. Mirny Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophysical Journal.*, 2004, 87:4021-4035.

[5] S. Ghosh, P. Ghosh, K. Basu, S. Das and S. Daefler. iSimBioSys: A Discrete Event Simulation Platform for 'in silico' Study of Biological Systems *Proceedings of 39th IEEE Annual Simulation Symposium*, 2006, AL, USA.

[6] J. Hasty and J. J. Collins. Translating the Noise. *Nature, Genet.*, 2002, 31, 13-14.

[7] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 1977, 81(25):2340-2361.

[8] H. Kitano. Cell Designer: A modeling tool of biochemical networks. *online at*, http://www.celldesigner.org/

[9] D. Adalsteinsson, D. McMillen and T. C. Elston. Biochemical Network Stochastic Simulator (BioNets): software for stochastic modeling of biochemical networks. *BMC Bioinformatics.*, March 2004.

[10] C.J. Morton-Firth and D. Bray. Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.*, 1998, 192: 117-28.

[11] Cell Illustrator. *online at*, http://www.fqspl.com.pl/

[12] C.T. Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004, 431:99-104.

[13] P. Ghosh, S. Ghosh, K. Basu, S. Das and S. Daefler. An Analytical Model to Estimate the time taken for Cytoplasmic Reactions for Stochastic Simulation of Complex Biological Systems. *2nd IEEE Granular Computing Conference*, May 10-12, 2006, Atlanta, USA.

[14] P. Ghosh, S. Ghosh, K. Basu, S. Das and S. Daefler. Modeling the diffusion process in the PhoPQ signal transduction system: A stochastic event based simulation framework. *Intl. Symp. on Computational Biology & Bioinformatics (ISBB)*, 2006.

[15] P. Ghosh, S. Ghosh, K. Basu, S. Das and S. Daefler. A stochastic model to estimate the time taken for Protein-Ligand Docking. *2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Sep. 2006, Canada.

[16] P. Ghosh, S. Ghosh, K. Basu, S. Das and S. Daefler. Stochastic Modeling of Cytoplasmic Reactions in Complex Biological Systems. *6th IEE International Conference on Computational Science and its Applications (ICCSA)*, May 8-11,