Modeling transcription factor binding events to DNA using a random walker/jumper representation on a 1D/2D lattice with different affinity sites

Vahid Rezania1,2, Jack Tuszyński1 and Michael Hendzel1

1 Division of Experimental Oncology, Cross Cancer Institute, 11560 University Avenue, Edmonton, AB T6G 1Z2, Canada
2 Institute for Advanced Studies in Basic Sciences, Zanjan 45195, Iran
E-mail: vrezania@phys.ualberta.ca

Received 1 October 2007
Accepted for publication 15 October 2007
Published 21 November 2007
Online at stacks.iop.org/PhysBio/4/256

Abstract
Surviving in a diverse environment requires corresponding organism responses. At the cellular level, such adjustment relies on the transcription factors (TFs) which must rapidly find their target sequences amidst a vast amount of non-relevant sequences on DNA molecules. Whether these transcription factors locate their target sites through a 1D or 3D pathway is still a matter of speculation. It has been suggested that the optimum search time is when the protein equally shares its search time between 1D and 3D diffusions. In this paper, we study the above problem using Monte Carlo simulations by considering a simple physical model. A 1D strip, representing a DNA, with a number of low affinity sites, corresponding to non-target sites, and high affinity sites, corresponding to target sites, is considered and later extended to a 2D strip. We study the 1D and 3D exploration pathways, and combinations thereof by considering three different types of molecules: a walker that randomly walks along the strip with no dissociation; a jumper that represents dissociation and then re-association of a TF with the strip at a later time at a distant site; and a hopper that is similar to the jumper but it dissociates and then re-associates at a faster rate than the jumper. We analyze the final probability distribution of molecules for each case and find that TFs can locate their targets on the experimental time scale even if they spend only 15% of their search time diffusing freely in the solution. This agrees with recent experimental results obtained by Elf et al (2007 Science 316 1191) and is in contrast to previously reported theoretical predictions. Our results also agree with the experimental evidence for the role of chaperons and proteasomes in stabilizing and destabilizing TFs binding, respectively, during the regulation process. Therefore, the results of our manuscript can provide a refined theoretical framework for the process.

1. Introduction
The development of an adult animal from a single cell relies upon the ability of sequence-targeted DNA regulatory proteins to coordinate the expression of genes in a development- and tissue-specific manner. These sequence-specific DNA binding transcription factors (TFs) must locate target sequences amidst a vast number of non-relevant sequences. Surprisingly, the binding processes to specific sites happen at very fast rates [2], approximately 100–1000 times faster than the upper limit of diffusion-controlled motion of molecules of the same size [3, 4]. The mechanism(s) whereby these regulatory proteins find their target sites in long DNA molecules in such a rapid way has been the subject of extensive theoretical and experimental investigations. Yet it is currently still a matter
of speculation. See review by Halford and Szczelkun [5] and Halford and Marko [6].

In one scheme known as a ‘sliding’ or ‘scanning’ mechanism, the DNA binding protein binds randomly at any site on the DNA and then translocates along the sequence until it finds its target [7, 4]. In this scenario proteins move along the DNA by executing 1D diffusion-controlled motion without losing their contacts with the DNA. At each time step the protein may move forward or backward (with equal probability) by taking one step along the DNA. This is equivalent to the symmetrical random walk problem in 1D.

The second mechanism, however, involves random walks between disassociation and re-association events in solution. In this mechanism two different modes of behavior are available to proteins, namely the so-called hopping and jumping mechanisms, whereby proteins dissociate from the DNA and move through the solution by 3D diffusion-controlled motion before they can re-associate with another site on the same DNA [8, 9]. Hopping refers to the case when the re-association occurs at a distance shallower than 20 bp while jumping implies re-association at >20 bp from the dissociated site. Although hopping and jumping rates depend on the DNA topology (linear, folded, etc), the overall rate of hopping would be higher than that of jumping [5, 6]. This is due to the low diffusivity of these macromolecules in solution (diffusion constant \( \leq 10 \mu m^2 s^{-1} \)), so that the majority of re-associations would happen at or very near to the dissociated site [7].

For those proteins with two or more binding sites (such as the Lac repressor, the SfiI and NgoMIV endonucleases) on a single DNA molecule, ‘interssegmental transfer’ is considered to be the third way that proteins move from one site to another. In this mechanism, a bound protein transiently binds to another site at the same time. After releasing, the protein may either remain at its initial binding site or move to a new site [10]. This mechanism, however, requires the juxtaposition of two sites within an interval shorter than the protein–DNA reaction radius in 3D space. It is shown that the interssegmental transfer mechanism is more likely for those proteins with two binding sites that are separated form each other by >300 bp [11, 12].

The sliding mechanism has been more attractive because of the assumption that diffusion along the DNA length is more rapid than a 3D search [2, 4, 7, 13]. Several experimental strategies have been developed to determine whether or not the facilitated diffusion in 1D is the dominant mechanism. In one procedure, a group of DNA molecules of different lengths, where each has a copy of the target site, is considered [4, 14–17]. The results, however, did not rule out 3D diffusion (e.g. hopping or jumping) and simply showed that the protein binding association rate decreases as the DNA molecule gets shorter. This can be explained by noting that the binding protein has less chance to encounter a shorter DNA fragment [5].

In an alternative attempt, Terry et al [18] studied the binding of the EcoRI restriction enzyme on both linear and circular 388 bp DNA molecules (fixed length molecules) with two target sites 51 bp apart. Interestingly, the protein binding strength depends on whether the two target sites are located closer to one end of DNA molecules or placed equidistant from both ends. In the former case, the enzyme showed more frequent cleavages at the innermost target site while in the latter case no preference between the target sites was observed. Binding to the circular DNA also showed no preference between the target sites; however, the degree of processivity was approximately two times higher than that on the linear DNA. Processivity is referred to as the fraction of total reactions that are cleaved at both target sites during a single binding event. Strikingly, the processivity results show that translocation through the hopping (3D) mechanism must be more efficient than the sliding (1D) mechanism, otherwise the degree of processivity would be marginally above that on the linear DNA [9].

Combining the above strategies, Stanford and Szczelkun [9] measured the processivity factor for a group of DNA molecules with the same length but each consisted of two EcoRV target sites where their inter-site spacing varied from one DNA to another ranging from 54 to 764 bp. Theoretically, by increasing the inter-site spacing, \( n \), the processivity factor decreases with a rate of \( n^2 \) for 1D diffusion and \( n^{3/2} \) for 3D diffusion [19]. Interestingly, the resulting processivity factor did not match the 1D diffusion relationship (being off by several orders of magnitude) but it approximately agreed with the 3D diffusion mechanism.

Halford and Szczelkun [5] also reviewed other experimental studies that were designed to determine whether a 1D or a 3D search is preferred by binding proteins and concluded that binding proteins find their target sites, both in vitro and in vivo, primarily through the 3D diffusion (hopping or jumping). This is, of course, in contrast to the perception that 1D diffusion along DNA molecules is the predominant pathway in locating target sites.

Several theoretical investigations, however, suggested that the optimum search time for a TF to locate its target site on the DNA strip is when the protein spends an equal amount of time sliding along the DNA and diffusing freely in the solution [8, 20]. A more precise calculation was carried out by Coppey et al [21] who proposed a stochastic model for the protein–DNA interaction by considering a series of one-dimensional diffusions of a single protein interrupted by several random jumps from site to site. They showed that for a DNA molecule with an intermediate length (~10–100 bp) the protein should spend a little bit more time (about 10% more) in the 1D pathway in order to minimize the search time. For a long DNA molecule (>500 bp), however, they recovered the result by [8, 20].

More recently Elf et al [1] performed an experimental investigation on lac repressor (LacI) in a living Escherichia coli cell and directly observed specific binding of the labeled lac repressor. Interestingly, using the single-molecule tracking technique, they determined 1D and 3D diffusion constants of dimeric LacI-Venus to be \( D_1 = 0.046 \pm 0.01 \mu m^2 s^{-1} \) (measured in vitro) and \( D_1 = 3 \pm 0.3 \mu m^2 s^{-1} \) (measured in vivo and without its DNA binding domain), respectively. However, using the mean-square displacement plot, they found that the effective diffusion constant of LacI in vitro is \( D_{eff} = 0.4 \pm 0.02 \mu m^2 s^{-1} \) that is ten times bigger than 1D constant. From this fact they concluded that TFs spend 87% of
the time sliding along DNA. This is obviously in contrast with the theoretical results found in [8, 20, 21]. Furthermore, they measured the residence time of TFs on non-specific sites to be about 0.3–5 ms and the searching time for a single TF to find a target site is estimated to be about 65–360 s. The residence time corresponds to the probability of TF staying bound to a particular site.

The above result more or less agrees with a single molecule atomic force spectroscopy experiment performed by Pant et al [22, 23] and Sokolov et al [24] who studied the relation of the protein binding rate with the protein concentration, C, in the solution. They argued that the rate of binding to a specific site is proportional to $C^2$ and $C$ for pure 1D and 3D diffusions, respectively. By considering bacteriophage T4 gene 32 protein, Pant et al [22, 23] found that at low to moderate protein concentration the binding rate to the specific site of a single-stranded DNA is more or less proportional to $C^2$. This shows that most of the time the protein is bound to the DNA.

In this study, we perform a set of numerical simulations intended to represent the interaction of binding proteins with DNA molecules. Our aim is to elucidate through numerical modeling whether a 1D or a 3D pathway is more favorable and to characterize the type of results that can be expected from a combination mode of behavior. Here, the 1D pathway refers to the case that molecules move along the lattice without dissociation from the lattice while the 3D pathway refers to the case that molecules dissociate and then re-associate as they move from site to site. We consider a binding protein to be a random walker (hopper/jumper) that walks (hops/jumps) along a one- or two-dimensional lattice representing DNA molecules. It is worth emphasizing that the walker and jumper represent two distinct modes of motion. When a binding protein is sliding along the DNA we identify it as a walker. When the same protein dissociates from the DNA and moves in the surrounding solution, we identify it as a jumper. In general, we specify that a molecule spends $q\%$ of its time in the walking mode and $(1-q) \times 100$ in the jumping mode, where $0 < q < 1$. However, in general some molecules may be in the walking mode and some may in the jumping mode. We also represent the TFs as ensemble in order to be able to perform statistical averaging with meaningful outcomes. Binding to a given DNA site depends on the binding energy between the molecule and the site which is quantified here through an affinity to the site. The smaller the affinity the bound molecule has the greater the chance to translocate to another site. In general, each site may have a different affinity from others. In this study, however, for simplicity and clarity of analysis, we assume the presence of only two different affinities resulting in low affinity (LA) sites and high affinity (HA) sites, respectively. The generalization to several non-equal affinity sites is straightforward and can be pursued in the future once a set of realistic binding energy values has been experimentally determined. The HA sites can be considered to be target (specific) sites that a binding protein is searching for. The LA sites are those non-specific sites that a protein moves along to find its eventual target site. Furthermore, we assume that each site can have up to a single molecule bound to it at a given time.

Based on the results of several experimental studies we believe that the model considered here, though still quite simplistic in its lack of detail, offers a realistic approximation for prokaryote (and possibly eukaryote) cells [1, 22–24].

2. Method

To represent the DNA substrate for TF binding, consider a one-dimensional lattice made of $L$ sites (in total) comprised of low affinity (LA) and high affinity (HA) sites. Such a system can be achieved, for example, by pulling DNA with optical tweezers to avoid DNA folding [22–24]. A molecule, representing a TF, may leave an LA site with a greater probability than an HA site. As a result, the probability of finding a molecule in an HA site is greater than that in an LA site. In terms of the binding energies, these sites can be represented by several potential wells placed in sequence along the DNA lattice in which HA sites have deeper potentials [20] (see figure 1). Recently, Long et al [26] studied the detection and prediction of the transcription factor binding sites in the proximal-promoter region of function-specific human genes and found that about 0.01–0.06% of the total LA sites are the HA sites. It is too computationally intensive to examine an abundance that is as low as the actual abundance. Here, as a ballpark figure we assume that 10% of the total number of sites accessible to TF proteins are HA sites distributed among LA sites in an a priori unknown fashion. It is worth noting that the absolute abundance does not influence the overall result of current study. To gain further confidence in our model we also ran simulations with lower population of high affinity sites (about 0.04%) and found very similar results. See figure 5 for detail.

The probability of leaving from an LA or HA site can be estimated by observing the residence time of a binding protein at non-specific or specific sites. Using the technique of fluorescent recovery after photobleaching (FRAP), the mean residence times for several chromatin-binding proteins have been studied [27–29]. The reader is referred to the review by van Holde and Zlatanova [30] for further references. Phair et al [29] studied over 20 chromatin proteins and distinguished
two slow and fast recovery populations with mean residence times ranging from ~3 to 6 s for the fast population and from ~15 to 30 s for the slow population. They found that Jun and XBP proteins have the shortest mean residence time of ~2 s and H10 has the longest one that could be up to ~3 min. See tables 1–3 in the paper by Phair et al [29]. Elf et al [1] also measured the residence time in non-specific sites to be about 5 ms. Furthermore, several experimental studies have shown that TFs do not remain permanently bound to a target site but rather undergo cycles of binding and unbinding [31, 32]. In particular, studying the exchange of the glucocorticoid receptor (GR) at promoter target sites has shown that while chaperon hsp90 enhances and then stabilizes the binding of the GR to the target site, the proteasome p23 induces the GR removal from the target site [32]. As a result, GR cycles between binding and unbinding with a time scale from ~1 min to ~1 h. This is due to the competitive role of chaperons and proteasomes present near each site during the transcription regulation process. While chaperons enhance the binding of TFs to target sites, the proteasomes degrade it [32]. We modeled this behavior by introducing specific sites with non-zero but small leaving probabilities. Here, we assumed that the molecule has a 20% chance to leave an HA site while it has a 67% chance to leave LA sites. We assume that the probability of leaving a particular binding site, $P$, is related to the binding energy, $E$, via an Arrhenius relationship: $P = P_0 \exp(-\Delta E/kT)$ as is expected for any chemical reaction with an activation energy barrier such as an enzymatic process. Here $k$ is the Boltzmann’s constant and $T$ is the temperature. Hence, the choice of the probability values for LA and HA sites corresponds to an energy difference between them of approximately $1.2kT$ which is at the low end of the expected levels. Energy differences of less than $kT$ would not result in a high- and low-affinity designation.

As we mentioned earlier, a molecule can walk or hop/jump along the lattice. A walker moves one step at a time without dissociating from the lattice. The walker moves left or right with an equal probability (symmetric walker). In the case of jumping, however, one molecule dissociates from a site (LA or HA) and after a random time interval, the same or another molecule associates at any available site (chosen randomly) along the lattice. The hopper, in contrast to a jumper, re-associates to the lattice instantly (if the free site is available) within a 100-site radius from the dissociated site. As a result, the hopper does not travel too far from the dissociated site. For the hopper/jumper there is no left or right preference. In the appendix, we have given some essential background information and mathematical formulae for the probability distributions for random walkers and random jumpers. The reason for our choice of a random re-association site is due to the typical range of values for the 3D diffusion constant which is on the order of $1 \, \mu\text{m}^2 \, \text{s}^{-1}$. Together with the formula $(x^2) = 6Dt$ and taking the value of $t$ on the order of 1 s gives the rms distance traveled in the surrounding medium on the order of 2–3 $\mu$m which indicates that the freed TF can find itself in an arbitrary location along the DNA upon re-binding.

The simulation starts from an initial state made of $M$ ($<L$) molecules that are randomly distributed along the lattice. For the walking case, a random number $r$ is chosen and then compared with the site’s probability, $P_0 = (P_{HA} \, \text{or} \, P_{LA})$. The site’s probability is plotted schematically in figure 1. As seen, 100 HA sites are clustered in the middle of the lattice with a total $L = 1000$ sites. If $r > P_0$ and its immediate neighbor is free, the walker will leave, otherwise it will stay. The immediate left or right neighbor is also chosen randomly using a separate uniform random distribution. Note that here we do not assume that the molecule is trapped in the HA for good. This assumption can change the final result drastically. We will go back to this point later. Similarly, in the case of hopping/jumping, a random number $r$ is chosen and then compared with the site’s probability, $P_0 = (P_{HA} \, \text{or} \, P_{LA})$. In this case only, a pool of molecules is also considered such that a new molecule from the pool may associate to a free site on the lattice and a molecule may dissociate from the lattice and enter the pool. The jumper will leave the site when $r > P_0$ and enter the pool of free TFs, otherwise it will stay bound to the DNA. After a random time interval, a destination site is chosen randomly using a separate uniform random distribution. If the new site is free, a molecule from the pool will associate with the DNA. In the case of hopping, the new destination site is chosen within a 100-site radius from the dissociated site with a diminishing probability with a distance. As a result, the adjacent sites to the disassociated site are more probable. The above procedure is repeated $N_{\text{sim}}$ times.

To represent the problem more realistically, we also studied the dynamics of TFs on a 2D lattice consisting of several affinity sites. In this case, in addition to LA and HA sites, a zero affinity site representing solution (water) is also considered (see figure 2). As seen in figure 2, as well as individual HA sites, there are some clustered HA sites along the $x$- and $y$-directions. Here, the $x$-direction refers to the motion along a particular DNA molecule and the $y$-direction refers to the motion along different DNA molecules in the immediate vicinity. Therefore, in this case TFs cannot only move along one DNA (similarly to a 1D strip), but they...
In total, over the course of our simulation each molecule, a walker or a jumper, will have $N_{\text{step}}$ (successful or unsuccessful) movement events. Note that each step represents a unit of time in the simulation. In general, we are interested in the total time (or number of steps) that it takes the system to reach the equilibrium. It is clear that the equilibrium distribution occurs when all the HA sites are filled. To have a better statistic, the simulation runs for a given number of steps (10, 100, 1000) and 10 000 steps (star). The HA sites are clustered in the middle.

The above results can be explained by noting that the pure jumping case is equivalent to a system with high protein concentration. So there are enough TFs around the DNA to bind the target sites immediately. In contrast, in the case of can also jump from one DNA to another. The clustered HA sites along the y-direction represent individual HA sites from different DNA molecules. This type of clustering, in fact, has been observed experimentally [33, 34] and it offers additional support to our model.

### 3. Results

In total, over the course of our simulation each molecule, a walker or a jumper, will have $N_{\text{step}}$ (successful or unsuccessful) movement events. Note that each step represents a unit of time in the simulation. In general, we are interested in the total time (or number of steps) that it takes the system to reach the equilibrium. It is clear that the equilibrium distribution occurs when all the HA sites are filled. To have a better statistic, the simulation runs for $N_{\text{sim}}$ times. Here, the final state for walkers (jumpers) is found after $N_{\text{step}} = 1000$, 10 000, 50 000 and 100 000 steps (10, 100 and 1000), each with $N_{\text{sim}} = 1000$. The results are shown in figures 3–9.

In figures 3 and 4 we assume that the 100 HA sites are clustered in the middle. Figure 3 demonstrates the final distribution of the 105 walkers after $N_{\text{step}} = 1000$ steps (blue circle), 10 000 steps (green diamond), 50 000 steps (red asterisk) and 100 000 steps (black star). The distribution is fairly uniform everywhere except close to the HA sites. The distribution decreases near the edges, suddenly increases at the edges and then decreases as it goes toward the middle. Interestingly, not all the HA sites have the highest probability of being occupied, except those who are placed at or near the edges. Figure 4 presents the probability of finding a randomly distributed ensemble of 105 hopper/jumpers after $N_{\text{step}} = 10$ steps (blue circle), 100 steps (red asterisk), 1000 steps (green diamond) in the lattice. It is clear that the equilibrium distribution is obtained after $N_{\text{jump}} \sim 100$ jumps. As expected, all the HA sites have the highest probability which is interestingly (more or less) constant for all curves. The LA sites, however, are shown to have more or less similar probabilities for each run with a limiting value $\sim 0.33 = 1 - P_0$. It is worth noting that the walker shows very different asymptotic behaviors in comparison with jumpers. It is clear that the longer the molecules walk or jump, the closer the final distribution approximates an equilibrium distribution as it should be $P_{\text{walk}}(x) \sim P_{\text{equilibrium}}(x) = 1 - P_0(x)$ [35] which can be seen as an additional verification of our model. Comparing figures 3 and 4 one can say that the hoppers/jumpers reach the final distribution in a much shorter time (or alternatively in a smaller number of steps) than walkers. As is readily seen, even after $N_{\text{walk}} = 100$ 000 steps, the final distribution of walkers does not match the equilibrium distribution completely. More interestingly, the square root relation is also observed between the required number of jump events and the number of walk events in order to reach the equilibrium distribution, i.e. $N_{\text{jump}} \sim \sqrt{N_{\text{walk}}}$. This is in agreement with experimental observations [9].

Figure 5 demonstrates the result of the simulation for low population of HA sites. Here we consider only four HA sites that are clustered in the middle of 10 000 LA sites. The results are obtained for $N_{\text{step}} = 100$ and $N_{\text{sim}} = 100$ 000. It is clear that the overall behavior of the probability distribution is very similar to figures 3 and 4 where the population of the HA sites is 10%. Therefore, we would not expect that the absolute abundance of the HA sites changes the overall result of current study. However, to get an adequate result we had to run the simulation for longer time (at least 100 times longer than figures 3 and 4).

The difference between the 1D (walking) and 3D (jumping) pathways is also studied for different LA and HA site distributions. Figure 6 demonstrates the final distributions for walkers and jumpers when 50 single HA sites are distributed among LA sites. In this case, none of the HA sites is clustered. Figure 7 represents the results for randomly distributed sites with different depths. Although in all cases the jumpers reach the equilibrium distribution on a shorter time scale, for a single HA site distribution (figure 6) the walkers reach the equilibrium distribution faster in comparison with the cases when the HA sites are clustered.

The above results can be explained by noting that the pure jumping case is equivalent to a system with high protein concentration. So there are enough TFs around the DNA to bind the target sites immediately. In contrast, in the case of...
Figure 5. Probability distribution walkers and jumpers after 100 steps with a low population of HA sites. Only four HA sites are clustered in the middle of 10,000 LA sites. The whole lattice is plotted in the top panel, while the middle 100 sites are shown in the bottom panel for comparison. Asterisk: walkers; circle: jumpers.

Figure 6. (a) Initial distribution of the 50 single HA sites (not clustered) with different depths. (b) Probability distribution of 105 jumpers after jumping for 100 steps for the initial distribution in (a). Probability distribution of 105 walkers after walking for (c) 100 steps and (d) 10,000 steps for the distribution in (a).

pure walking, the protein concentration in the surrounding solution is virtually zero. Therefore, TFs will find their target sites after scanning more or less the entire lattice. To be more realistic, we consider a mixed case. Starting from the case of 100% walkers, 0% jumpers, we increase the jumpers’ percentage in the system (i.e. molecules can now either walk or jump). Interestingly, we find that having reached the mixture of 85% walkers, 15% jumpers, the equilibrium distribution can be achieved after $N_{\text{step}} \sim 100–1000$ steps. This agrees with the experimental results found by Elf et al [1] and contrasts with the theoretical expectation [8, 20, 21]. In figure 8, we demonstrate the result of three different mixtures of random walkers and jumpers, (95%, 5%), (90%, 10%) and (85%, 15%) after $N_{\text{step}} = 100$. As a result, if TFs spend about 15% of their time diffusing freely in the solution, they can find their target sites in a very short time.

In the above simulations, we assumed that the molecules can escape from the HA sites (though, with a smaller probability) and so they would not be trapped in the HA sites for good after binding to them. In order to see how this
assumption would affect the results, we ran some simulations by assuming that the TF molecules will be trapped for good after entering HA sites. Interestingly, we find that the equilibrium distribution cannot be achieved in a short time (100–1000 steps) unless more than 80% of the molecules are jumpers. So in this case the protein should be unbound for most of the time, which is in contrast with experimental results [1]. Interestingly, such a continuous association–dissociation process has been experimentally observed for the glucocorticoid receptor [31, 32] and supports our assumption.

It is generally known that a symmetric random walk distribution approaches the Gaussian distribution after a
Modeling transcription factor binding

Figure 9. $\langle |x|^3 \rangle / \langle x^2 \rangle^{3/2}$ value as a function of the number of steps for 105 walkers, jumpers, hoppers and/or mixed molecules. The 100 HA sites are clustered in the middle. Blue line-circle: walker with $N_{\text{step}} = 5000$; red line-circle: walker with $N_{\text{step}} = 10000$; black line-circle: walker with $N_{\text{step}} = 50000$; green line-circle: walker with $N_{\text{step}} = 100000$; blue line-square: jumper with $N_{\text{step}} = 5000$; red line-square: jumper with $N_{\text{step}} = 10000$; blue line-triangle: hopper with $N_{\text{step}} = 5000$; black line-triangle: mixed (50% walker, 25% jumper, 25% hopper) with $N_{\text{step}} = 5000$; purple line-triangle: mixed (80% walker, 10% jumper, 10% hopper) with $N_{\text{step}} = 5000$. Those data with $N_{\text{step}} > 5000$ are scaled.

Figure 10. (a) Expected final distribution of molecules for probability given in figure 2. Probability distribution of 120 walkers after walking for (b) 100 steps, (c) 10000 steps and (d) 100000 steps.

sufficiently long run, i.e. $N_{\text{step}} \to \infty$. This is a consequence of the central limit theorem. In general, the required number of steps to achieve a desired level of convergence to the equilibrium distribution can be estimated using the Berry–Esséen theorem [36] as

$$N \geq \frac{25}{4} \frac{\langle |x|^3 \rangle^2}{\langle x^2 \rangle^3} \frac{1}{\epsilon^2},$$

where $\epsilon$ represents the level of convergence. It is clear that the criterion 1 relies on the ratio between the second, $\langle x^2 \rangle$, and third, $\langle |x|^3 \rangle$, moments of the random walk/jump probability distribution. In the appendix we calculate the ratio $\langle |x|^3 \rangle / \langle x^2 \rangle^{3/2}$ for the three different cases studied in this paper: a random walk probability $p_w(x)$, a random jump probability $p_j(x)$ and a mixed probability $p_{wj}(x)$, with results of 1.000, 1.2990 and 1.1482, respectively. Figure 9 represents the ratio $\langle |x|^3 \rangle / \langle x^2 \rangle^{3/2}$ as a function of $N_{\text{step}}$ for 105 walkers, jumpers, hoppers and/or mixed for the initial clustered distribution (see figure 1). The results for all runs are in the range of $1.3 \pm .03$ which is close to the random jump distribution.

As mentioned earlier, in order to make the model more realistic, we also performed several simulations with a 2D lattice (figure 2). In general, one would expect several DNA molecules to be present in the solution at nearby locations. As a result, TFs that are moving in the bulk have now more chances to bind. Similarly to a 1D lattice, we consider two different types of molecules: walkers and hoppers/jumpers. We assume that a walker can only walk along the DNA (the x-direction) except when it encounters
zero affinity sites. In those sites, the walker can move in both the x- and y-directions. Hoppers/jumpers can move in both directions with no restriction. We used similar strategy for walking and hopping/jumping from one site to another as we employed in the 1D lattice (see above). Results are shown in figures 10–12.

Figure 10 shows the probability distribution of 120 walkers after walking for 100, 10 000 and 100 000 steps. To compare, we also show the expected final distribution of molecules $P_{\text{equilibrium}}(x, y) \sim 1 - P_0(x, y)$ for the probability given in figure 2. The probability distribution of 120 jumpers after jumping for 100 steps is shown in figure 11. Comparing this with the expected distribution, it is clear that the jumpers reach the equilibrium distribution in a shorter time (smaller steps) than walkers.

Figure 12 demonstrates the probability distribution of 120 mixed walkers and jumpers after walking/jumping for 1000 steps. As can be readily seen, by increasing the number
of jumpers, the final distribution rapidly approaches to the equilibrium distribution. This is again in good agreement with the experimental results obtained by Elf et al [1].

4. Discussion and conclusions

It is well known that very complicated processes such as the coordination of gene expression, cellular metabolism, and organ and tissue development rely on a more fundamental process, the regulation of gene transcription through transcription factor binding. This process coordinates the expression of genes in a tissue-specific manner during early stages of development and tissue specification and is crucial in the development of an adult animal from a single cell. The fundamental question is how do these transcription factors find their target sites amongst a large number of non-target sites. This is still an unresolved issue. Specifically, experimental results suggest that transcription factors must locate target sequences at very fast rates [2], approximately 100–1000 times faster than the upper limit of a diffusion-controlled motion of molecules with the same dimension [3, 4]. An enormous amount of effort has been spent trying to understand whether transcription factors locate their targets through a 1D (sliding along the DNA) or 3D (diffusion in solution) pathway. Although earlier investigations favored the 1D pathway, recent experimental studies suggest that the mixture of 1D and 3D pathways is more efficient [1].

In this paper, we have addressed the above question using a set of Monte Carlo simulations involving a 1D strip with a number of potential wells at different depths that are clustered on the strip. The strip represents a DNA molecule and each potential well represents a binding site. In the simplest case we distinguish two potential depths: low affinity (LA) sites with a shallow potential and high affinity (HA) sites with a deep potential (see figure 1). The generalization to several potential depths is straightforward.

The strip is located in a pool of molecules (a 2D lattice with zero affinity sites) that can associate with or dissociate from it. In our simulations, however, we have considered three different types of molecules: a walker that performs a random walk along the strip without dissociation from it; a jumper that dissociates from the strip and then re-associates with the strip at a distant site after a random time interval; and a hopper that dissociates and then re-associates at a faster rate than the jumper. As a result, the hopper travels less distance than the jumper before re-association. The walker represents a binding protein that slides along the DNA to find its target site via the 1D pathway. The jumper represents a binding protein that slides along the DNA to find its target site via the 1D pathway. We consider a 2D lattice rather than a simple 1D lattice in our simulations, however, we have considered three different types of molecules: a walker that performs a random walk along the strip without dissociation from it; a jumper that dissociates from the strip and then re-associates with the strip at a distant site after a random time interval; and a hopper that dissociates and then re-associates at a faster rate than the jumper. As a result, the hopper travels less distance than the jumper before re-association. The walker represents a binding protein that slides along the DNA to find its target site via the 1D pathway. The jumper represents a binding protein that slides along the DNA to find its target site via the 1D pathway. Reassuringly, our results also agree with the experimental data reported in [31, 32] on the role of chaperones and proteasomes during the TF regulation process.

In recent years, however, some experimental investigations suggest a new role for chromatin in transcription factor regulation. In particular, studying the exchange of the glucocorticoid receptor (GR) at promoter target sites has shown that while the chaperone hsp90 enhances and then stabilizes the binding of the GR to the target site, the proteasome p23 induces the GR removal from the target site [32]. As a result, GR undergoes cycles of binding and unbinding on a time scale from ~1 min to ~1 h and not binding permanently to the site. This confirms our assumption that the TF can leave specific sites with a non-zero but small leaving probability. This is again in contrast with previous theoretical investigations that assumed the TF will be trapped forever after entering the target site.

Several simulations with different molecules and a different number of steps have been performed: walkers only with \( N_{\text{step}} = 1000, 10,000, 50,000 \) and 100,000; jumpers/hoppers only with \( N_{\text{step}} = 10, 100 \) and 1000; and a mixture of walkers, jumpers and hoppers. Our results are plotted in figures 3–9. As seen in figures 3 and 4, jumpers reach the equilibrium distribution having taken on the order of \( N_{\text{jump}} \sim 100 \) steps while the final distribution of walkers even after \( N_{\text{walk}} = 100,000 \) steps is quite different from the expected equilibrium distribution. Interestingly, we find that \( N_{\text{jump}} \sim \sqrt{N_{\text{walk}}} \), which is in good agreement with experimental observations [9]. In the case of mixed molecules, we have examined different populations of molecules and found that even if only 15% of the molecule population are not walkers, the results are very similar to the jumper case. See figure 8 that shows the results of three different mixtures of random walkers and jumpers, (95%, 5%), (90%, 10%) and (85%, 15%) after \( N_{\text{step}} = 100 \). Our results indicate that if the TFs spend just about 15% of their search time diffusing in the solution they are able to locate their target site in a short period of time. This indeed agrees with the experimental results obtained in [22–24] and specifically by Elf et al [1] who found 87% of the time is spent by TFs diffusing along the DNA. However, our results are in contrast with the theoretical expectation that suggests the optimum search time happens when the protein spends equal time sliding along the DNA and diffusing freely in the solution (a 50–50 mixture) [8, 20, 21].

As expected, similar results are also obtained when we consider a 2D lattice rather than a simple 1D lattice (figure 2). It is more realistic to assume that several DNA molecules are present in the solution at different locations. As a result, TFs that are moving in the bulk have now more chances to bind. The results of our simulations are shown in figures 10–12. We compared the probability distribution of 120 walkers after walking for 100, 10,000 and 100,000 steps with that of 120 jumpers after jumping for 100 steps as shown in figures 10 and 11. It is clear that the jumpers reach the equilibrium distribution in a shorter time (smaller steps) than walkers.

In recent years, however, some experimental investigations suggest a new role for chromatin in transcription factor regulation. Chromatin consists of a fundamental unit called the nucleosome, a disc-shaped octamer of eight histone proteins, bound to DNA. Each histone interacts with other histones and DNA to form the nucleosome. A chromatin-mediated mechanism may enhance fidelity of transcriptional regulation and control of gene expression in the cell. Initial experiments monitoring histone H1-GFP fusion during FRAP experiments have shown a rapid exchange of the H1 with sites (with few minutes residence time) the chromatin remains positionally stable [27, 28]. A number of chromatin-binding proteins have...
subsequently been monitored by Phair et al [29] in vivo using FRAP method. They found that transient binding is a common property among chromatin-associated proteins. They also demonstrated that all of these chromatin-binding proteins show rapid mobility. They concluded that these proteins continuously scan the genome space for appropriate binding sites through diffusional hopping (the 3D pathway) between chromatin fibers. Buck and Leib [38] also studied repressor–activator protein 1 in the yeast cells and found that these proteins locate their binding sites via a newly dynamic target specification mechanism. Chromatin-mediated regulation of accessibility coordinates genomewide distribution of DNA sequence motifs to target sites by remodeling the genome itself.

Future outlook

The present paper can be further extended by applying our methodology to more realistic potentials with a diverse distribution of HA and LA sites whose binding energy values are determined empirically. In fact, in a not-too-distant future it should be possible to obtain a binding energy landscape for DNA sequences of interest from molecular dynamics simulations performed for various TF proteins. This type of calculation could also incorporate the effects of solvent ions, pH and temperature on the effective binding potential due to electrostatic effects. It would be interesting to compare various DNA sequences and different TF molecules to find out how specific the mechanisms of molecular recognition are.

Acknowledgments

This research was supported by grants from the ACB, NSERC, CSA and CIHR. Support from the Allard Foundation is gratefully acknowledged by JAT. MJH is an AHFMR senior scholar. We thank Romain Sibi for his assistance with some of our numerical simulations.

Appendix. Calculating \( \langle |x|^3 \rangle / \langle x^2 \rangle^{3/2} \) for different distribution functions

The criteria 1 can be found as

\[
\frac{\langle |x|^3 \rangle}{\langle x^2 \rangle^{3/2}} = \begin{cases} 
1 & \text{for random walk distribution,} \\
1.299 \frac{(\alpha/4)(L/2)^3 + \beta \ell^3}{[(\alpha/3)(L/2)^2 + \beta \ell^2]^{3/2}} & \text{for random jump distribution,} \\
\frac{(\alpha/3)(L/2)^3 + \beta \ell^3}{[(\alpha/3)(L/2)^2 + \beta \ell^2]^{3/2}} & \text{for random walk–jump distribution,}
\end{cases}
\]  

(A.1)

where for a distribution function \( p(x) \), \( \langle x^n \rangle \) and \( \langle |x|^n \rangle \) are defined as

\[
\langle x^n \rangle = \int_{-\infty}^{\infty} x^n p(x) \, dx \quad \text{and} \quad \langle |x|^n \rangle = \int_{-\infty}^{\infty} |x|^n p(x) \, dx.
\]

The distribution function for a non-equal step size random walk in the range \( [-\ell, \ell] \) is given by

\[
p_w(x) = a \delta(x + \ell) + b \delta(x - \ell),
\]

(A.2)

where \( \delta(x) \) is the Dirac delta function and \( a + b = 1 \). The distribution function for a random jump in the range \( [-L/2, L/2] \) can be written as

\[
p_j(x) = (1/L)(H(x + L/2) - H(x - L/2)),
\]

(A.3)

where \( H(x) \) is the Heaviside function:

\[
H(x) = \begin{cases} 
0 & x < 0, \\
1 & x \geq 0.
\end{cases}
\]

Finally, the distribution function for a mixed random walk–jump in the range \( [-L/2, L/2] \) where \( \ell \leq L/2 \) can be written as

\[
p_{w+j}(x) = \alpha p_w(x) + \beta p_j(x),
\]

\[
= \frac{\alpha}{L} (H(x + L/2) - H(x - L/2)) + \beta (a \delta(x + \ell) + b \delta(x - \ell)),
\]

(A.4)

where \( \alpha + \beta = 1 \). For a special case \( \alpha = 1/2 = \beta \) and \( \ell = L/2 \), we find \( \langle |x|^3 \rangle/\langle x^2 \rangle^{3/2} \approx 1.1482 \).

References

[1] Elf J, Li G and Xie X S 2007 Probing transcription factor dynamics at the single-molecule level in a living cell Science 316 1191–4
[2] Riggs A D, Burgeoise S and Cohn M 1970 The lac repressor–operator interaction: 3. Kinetic studies J. Mol. Biol. 53 401–17
[3] Berg O G and von Hippel P H 1985 Diffusion-controlled macromolecular reactions Annu. Rev. Biophys. Chem. 14 131–60
[4] Shimamoto N 1999 One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements J. Biol. Chem. 274 15293–6
[5] Halford S E and Szczelkun M D 2002 How to get from A to B: strategies for analysing protein motion on DNA Eur. Biophys. J. 31 257–66
[6] Halford S E and Marko J F 2004 How do site-specific DNA-binding proteins find their targets? (Survey and Summary) Nucl. Acids Res. 32 3040–52
[7] von Hippel P H and Berg O G 1989 Facilitated target location in biological systems J. Biol. Chem. 264 675–8
[8] Berg O G, Winter R B and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: 1. Models and theory Biochemistry 20 6929–48
[9] Stanford N P, Szczelkun M D, Marko J F and Halford S E 2000 One- and three-dimensional pathways for proteins to reach specific DNA sites EMBO J. 19 6546–57
[10] Milsom S E, Halford S E, Embleton M L and Szczelkun M D 2001 Analysis of DNA looping interactions by type II restriction enzymes that require two copies of their recognition sites J. Mol. Biol. 311 515–27
[11] Podtelezhnikov A A and Vologodskii A V 2000 Dynamics of small loops in DNA molecules Macromolecules 33 2767–71
[12] Ringrose L, Chabanis S, Angrand P O, Woodroofe C and Stewart A F 1999 Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances EMBO J. 18 6630–41
[13] Richter P H and Eigen M 1974 Diffusion controlled reaction rates in spheroidal geometry. Application to
Modeling transcription factor binding

repressor–operator association and membrane bound enzymes

[14] Ehbrecht H J et al 1985 Linear diffusion of restriction endonucleases on DNA J. Biol. Chem. 260 6160–6

[15] Jack W E, Terry B J and Modrich P 1982 Involvement of outside DNA sequences in the major kinetic path by which EcoRI endonuclease locates and leaves its recognition sequence Proc. Natl Acad. Sci. USA 79 4010–4

[16] Jeltsch A and Pingoud A 1998 Kinetic characterization of linear diffusion of the restriction endonuclease Eco RV on DNA Biochemistry. 37 2160–9

[17] Jeltsch A, Wenz C, Stahl J and Pingoud A 1996 Linear diffusion of the restriction endonuclease Eco RV on DNA is essential for the in vivo function of the enzyme EMBO J. 15 5104–11

[18] Terry B J, Jack W E and Modrich P 1985 Facilitated diffusion during catalysis by Eco RI endonuclease. Nonspecific interactions in Eco RI catalysis J. Biol. Chem. 260 13130–7

[19] Doi M and Edwards S F 1984 The Theory of Polymer Dynamics (Oxford: Oxford University Press)

[20] Slutsky M and Minry L A 2004 Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential Biophys. J. 87 4021–35

[21] Coppey M, Benichou O, Voituriez R and Moreau M 2004 Kinetics of target site localization of a protein on DNA: a stochastic approach Biophys. J. 87 1640–9

[22] Pant K, Karpel R L, Rouzina I and Williams M C 2004 Mechanical measurement of single-molecule binding rates: kinetics of DNA Helix-destabilization by T4 gene 32 protein J. Mol. Biol. 336 851–70

[23] Pant K, Karpel R L, Rouzina I and Williams M C 2005 Salt dependent binding of T4 gene 32 protein to single and double-stranded DNA: single molecule force spectroscopy measurements J. Mol. Biol. 349 317–30

[24] Sokolov I M, Metzler R, Pant K and Williams M C 2005 Target search of N sliding proteins in a DNA Biophys. J. 89 895–902

[25] Robertson G et al 2007 Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing Nature Method 4 651

[26] Long F et al 2004 Genome-wide prediction and analysis of function-specific transcription factor binding sites In Silico Biol. 4 395–410

[27] Lever M A, Th'ng J P H, Sun X and Hendzel M J 2000 Rapid exchange of histone H1.1 on chromatin in living human cells Nature 408 873–6

[28] Misteli T, Gunjan A, Hock R, Bustin M and Brown D T 2000 Dynamic binding of histone H1.1 to chromatin in living cells Nature 408 877

[29] Phair R D et al 2004 Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins Mol. Cell. Biol. 24 6393

[30] van Holde K and Zlatanova J 2006 Scanning chromatin: a new paradigm? J. Biol. Chem. 281 12197

[31] Elbi C, Walker D A, Romero G, Sullivan W P, Toft D O, Hager G L and DeFranco D B 2004 Molecular chaperones function as steroid receptor nuclear mobility factors PANS 101 2876

[32] Stavreva D A, Müller W G, Hager G L, Smith C L and McNally J G 2004 Rapid glucocorticoid receptor exchange at a promoter is coupled to transcription and regulated by chaperones and proteasomes Mol. Cell. Biol. 24 2682

[33] Osborne C S et al 2004 Active genes dynamically colocalize to shared sites of ongoing transcription Nat. Genet. 36 1065

[34] Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M and van Steensel B 2006 Characterization of the Drosophila melanogaster genome at the nuclear lamina Nat. Genet. 38 1005

[35] van Kampen N G 1981 Stochastic Processes in Physics and Chemistry (Amsterdam: North-Holland) chapter VIII p 209

[36] Feller W 1996 An Introduction to Probability Theory and Its Application vol 2 (New York: Wiley) chapter XVI.8 p 525

[37] Santos-Rosa H et al 2002 Active genes are tri-methylated at K4 of histone H3 Nature 419 407–11

[38] Buck M J and Lieb J D 2006 A chromatin-mediated mechanism for specification of conditional transcription factor targets Nat. Genet. 38 1446–51