Effect of DNA conformation on facilitated diffusion

Chris A. Brackley¹, Mike E. Cates and Davide Marenduzzo
SUPA (Scottish Universities Physics Alliance), School of Physics and Astronomy, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, U.K.

Abstract
Within a living cell, site-specific DNA-binding proteins need to search the whole genome to find a target of ~10–20 bp. That they find the target, and do so quickly, is vital for the correct functioning of the DNA, and of the cell as a whole. The current understanding is that this search is performed via facilitated diffusion, i.e. by combining three-dimensional bulk diffusion within the cytoplasm or nucleoplasm, with one-dimensional diffusion along the DNA backbone, to which the protein binds non-specifically. After reviewing the standard theory of facilitated diffusion, we discuss in the present article the still rather rare direct computer simulations of this process, focusing on the three-dimensional part of the search, and the effect of DNA looping and the general DNA conformation on its efficiency. We close by highlighting some open questions in this field.

The search for a target on the DNA, and facilitated diffusion
In living cells, proteins routinely need to bind to a specific base pair sequence positioned on the DNA, i.e. they must find a small target on a large DNA molecule. For example, a transcription factor might bind to a promoter to initiate the transcription of one gene, or to suppress another. The size of such targets is often of the order of 10 bp, whereas the total size of the genome through which the protein needs to search is of the order of a few megabase pairs for bacteria, or a few billion base pairs for humans. A well-known case is that of the Escherichia coli lac repressor which binds to a target on λ-DNA; these binding dynamics have been well studied in vitro [1].

Importantly, the search has to be both accurate and fast: the protein should not become ‘stuck’ in a DNA region other than the target. One would expect that a transcription factor would inevitably interact strongly and non-specifically with the DNA, and this interaction would severely hinder its motion. However, recent in vivo experiments suggest that the whole search process is much faster than might be expected; for example, Elf et al. [2] measured a total search time of between 65 and 360 s for the lac repressor in E. coli, and a diffusion constant \( D \) for the lac repressor when diffusing freely in the cytoplasm of \( 3 \pm 0.3 \ \mu m^2 \cdot s^{-1} \). Given a cell volume \( V \) of \( \sim 2 \ \mu m^3 \) and a target of size \( a \sim 1 \ nm \), a crude estimate of the search time expected if the protein were to undergo only free three-dimensional diffusion is \( V/(Da) \sim 660 \ s \), longer than the measured time. It is remarkable that the search time can reach anywhere near the three-dimensional diffusion limit, given that we expect the protein to interact with the several megabase pairs of the bacterial genome.

How then do proteins reach their specific target region on the genome so quickly? The common explanation is that the proteins alternate between phases of free diffusion through the cytoplasm, and phases in which they move along the DNA, effectively performing one-dimensional diffusion along its backbone [2–6]. This one-dimensional diffusion is possible because the non-specific interaction the proteins have with the DNA has a smaller affinity than (and may be of a different nature to) the one which they have for their specific target. The exact mechanism of the one-dimensional motion is, however, still not well understood; initially it was thought that the electrostatic interaction between the protein and the negatively charged backbone of the DNA would allow the protein to effectively ‘slide’ along its contour [7]. A more recent suggestion [8–11] is that the protein has two different DNA-interaction modes, a weak one and a stronger ‘recognition mode’, the latter possibly involving part of the protein sitting in the major groove of the double helix, and leading to spiralling while sliding [12,13]. Despite the lack of a full description of the motion, measurements of the one-dimensional diffusion constant have been made in vitro [6], and the proportion of time spent in each of the one- and three-dimensional search modes has been measured in vivo [2].

The combined three-dimensional and one-dimensional diffusional dynamics is referred to as ‘facilitated diffusion’, and, as we shall see shortly, it can greatly reduce the mean search time. This is provided that the average length of DNA explored during each one-dimensional sliding episode (known as the ‘sliding length’) is chosen appropriately. Figure 1(A) shows the different transport mechanisms schematically. Direct in vivo experimental evidence of facilitated diffusion has recently been found in bacteria [14].

A scaling estimate of the mean search time
A simple and useful scaling argument to estimate the mean time, \( \tau_s \), that a protein takes to find a target on the DNA through facilitated diffusion was proposed by Halford and Marko [15], and we review this in the present article.
Proteins search for their targets via facilitated diffusion

When not interacting with the DNA, the protein performs three-dimensional diffusion; many authors differentiate between short-range excursions, where the protein returns to the DNA at a location near to where it left (sometimes known as ‘hops’), and long-range excursions, where the protein rejoins the DNA at a point far from where it left (sometimes called ‘long-range jumps’). Also, by virtue of DNA looping or other topological features, there may be correlations between the locations of successive one-dimensional sliding events. Finally, looping can also allow intersegmental transfers, where a protein with multiple DNA-binding sites can transfer directly between different DNA segments by transiently binding to both.

The key parameters on which this argument depends are the total DNA length $L$, the volume of the cell $V$, the three-dimensional and one-dimensional diffusion coefficients $D_3$ and $D_1$ respectively, and, finally, the sliding length $l_s$ mentioned above. To obtain an analytic expression for $\tau$, Halford and Marko [15] estimated the average duration of a three-dimensional excursion and that of a one-dimensional sliding event. Dimensional analysis suggest that the former is given by $\tau_{3D} \sim V/(D_3L)$, whereas the latter is $\tau_{1D} \sim l_s^2/D_1$. The mean number of search rounds (i.e. sliding events) can be approximated as $L/l_s$. This simplification hinges on the assumption that the point at which the protein hops off the DNA is uncorrelated with the point at which it lands after performing a three-dimensional excursion. It also neglects intersegmental transfers through which the protein exploits a loop on the DNA to jump from one segment to a distant one, without a three-dimensional excursion. Within this scaling theory, one can combine the expressions above to arrive at the following estimate for the mean search time:

$$\tau \sim AV/(D_3l_s) + BLI_s/D_1 \tag{1}$$

where $A$ and $B$ are dimensionless order unity prefactors which require simulations (or a more detailed calculation [3,16]) to be determined accurately; they cannot be deduced from the simple scaling theory outlined above.

Eqn (1) is arguably the most important prediction of the standard facilitated diffusion theory, and it is plotted in Figure 1(B) using parameters suitable for *E. coli*. Its remarkable consequence is that, because the dependence of the mean search time on the sliding length is non-monotonic, there exists an optimal value of $l_s$ which minimizes $\tau_s$. This optimal sliding length is $l_s^* = [(AD_1V)/(BD_3L)]^{1/3}$. If one assumes that $A = B$, and uses the values for the one- and three-dimensional diffusion constant measured in [2], then for a genome of 4.2 Mbp (with 0.34 nm/bp) and a volume of $2 \mu m^3$ (such parameters hold for *E. coli*), we obtain an optimal sliding length of approximately 4.6 nm.

Although the scaling theory leading to eqn (1) is simple and appealing, it is worth noting that it involves some quite significant approximations. As noted above, intersegmental transfers are disregarded; furthermore, the neglect of correlations between take-off and landing points on the genome effectively means that the DNA is considered as structureless. Another important simplification is that the theory does not consider the relaxation dynamics of the polymer [17], which is slow enough to cause chain structure to persist between successive episodes of one-dimensional exploration. This means that there is an underlying assumption that the polymer configuration changes on a time scale which is much shorter than the time scale of the search, which is not true in general.

How then, can one go beyond eqn (1) and refine the theoretical study of facilitated diffusion? One way is to resort to more complicated analytical theories. Unfortunately, such theories quickly become much less tractable than...
the arguments reviewed above. Another avenue, which we concentrate on in the present article, is to use computer simulations at different levels of coarse graining. Such simulations are still rather scarce to date, but have already helped to shed light on the physics of the search for a target along the genome. We now review such simulation studies in some detail.

**A review of numerical simulations of facilitated diffusion**

The first large-scale simulations of facilitated diffusion only date back to around 2006 [16,18,19], some 20 years after the seminal work by von Hippel and Berg [3] first proposed the concept that a combination of three-dimensional and one-dimensional diffusion can, potentially, speed up the search process. The outcome of the simulations by Klenin and colleagues [16,18,19] is, essentially, a quantitative verification of eqn (1). By combining a more accurate analytical theory and numerical simulations, the authors could estimate the values of the prefactors $A$ and $B$; the ratio $A/B$ was found, to a good approximation, to be equal to $1/(2\pi)$, with the individual values of $A$ and $B$ depending on the target size. The case considered by these first simulations was that of a semi-flexible and self-avoiding polymer, with parameters set so as to resemble DNA, confined inside an isotropic spherical container. The DNA conformation was kept fixed during the search process, thereby addressing the case in which the polymer relaxation is very slow with respect to the facilitated diffusion process (this is the opposite limit to that considered in the analytical treatment above). More recently, Foffano et al. [20] used a similar approach, but they considered the case of confinement within anisotropic geometries, such as cylinders, which may resemble more closely the case of bacterial cells. The authors performed dynamic Monte Carlo simulations and showed that, especially for large values of the aspect ratio of the confining geometry and short contour length of the DNA, the predictions from eqn (1) and the refinement in [16] do not give a good fit to the numerical data [20]. Furthermore, there did not appear to be a deep minimum of the mean search time curves, so that the idea that the sliding length must be precisely optimized in order to speed up the search appears questionable under those conditions. Another class of simulation neglects the DNA conformation completely, instead focusing on the one-dimensional dynamics and treating the three-dimensional motion in a mean field sense. Sokolov et al. [21,22] performed Monte Carlo simulations to find the mean first passage time of proteins moving on a homogeneous one-dimensional lattice which represents the DNA. They treated multiple proteins, and were able to confirm the results of their analytical continuum model. Slutsky and Mirny [9] considered a one-dimensional random walk in a sequence dependant potential, and Zabet and Adryan [23] extended this idea, using real DNA sequences and measurements of transcription factor DNA-binding affinities to construct a detailed sequence specific lattice model. They considered multiple proteins which can perform a random walk along the DNA with diffusion rates determined by the base pair sequence. The proteins can leave the DNA and move into a ‘reservoir’, returning to different sites of the DNA with rates chosen to give a mean field representation of the three-dimensional diffusion.

Sheinman and Kafri [24], who also concentrated on a lattice model, focused on the effect of intersegmental transfers, which are outside the scope of the scaling theory of eqn (1). The conclusion of this detailed study was that intersegmental transfers can make a remarkable difference to the quantitative estimate of the mean search time, increasing the efficiency of facilitated diffusion significantly. The simulations also suggested that there are several cases in which the search time depends only weakly on the sliding length (a conclusion similar to that of the later study by Foffano et al. [20]). The simulations also revealed a strong dependence on the DNA length, beyond that predicted by eqn (1).

All of the studies highlighted thus far considered the dynamics of the DNA in one of two limits: either the DNA conformation is quenched, i.e. frozen in time, or assumed to be changing so quickly that it need not be considered at all. Although, at first sight, these may seem to be reasonable approximations, such simplifications actually lead to quantitative differences, especially when intersegmental transfers are frequent [25,26]. Florescu and Joyeaux [27,28] chose instead a Brownian dynamics algorithm, where the DNA and protein interacted both sterically and electrostatically (à la Debye–Hückel), and the DNA was mobile. This interesting framework naturally takes into account both the DNA and the protein dynamics, and the time scales associated with each process are correctly captured. This model considered a coarse-grained (negatively charged) DNA (one bead corresponding to approximately $15$ bp), and a uniformly (positively) charged spherical protein. A higher level of detail can be included in Brownian dynamics studies [13,29], but, owing to the high computational overhead, only very short DNA molecules can be simulated in that case. The DNA of Florescu and Joyeaux [27,28] was confined in a small volume, and was fragmented into small strands so as to avoid tight turns (such turns would not appear in bacterial cells due to the larger cell volume, but crop up in simulations due to the smaller size used for computational efficiency). The most interesting result from these Brownian dynamics simulations was that the efficiency of facilitated diffusion should be smaller than discussed above, to the extent that the process could actually hinder target location for sufficiently strong DNA–protein interactions. In part, this was caused by a caging effect through which one protein could interact with several DNA segments, which then slowed down its diffusive motion.

A similar Brownian framework was used in our recent simulations [30]; however, we considered a more detailed ‘patchy sphere’ model for the protein. The interaction with the DNA was only via a small binding site on the protein surface, which avoids (or at least strongly reduces) the caging effect. Such a forcefield is a more suitable model.
Figure 2 | DNA conformation affects facilitated diffusion

(A) Sketch of the coarse-grained DNA and protein models used in [30]. The DNA is modelled as a bead and spring polymer with beads representing 2.5 nm of DNA (approximately 8 bp), with adjacent beads held together by springs; the chain is semi-flexible with a persistence length of 20 beads (approximately 50 nm; relevant for naked DNA). The protein is modelled as two spheres: one large sphere which interacts with the DNA via a steric repulsion only, and one smaller sphere, which represents the binding site of the protein. The latter sticks to the DNA with an interaction strength $\varepsilon$. The DNA is arranged as a sequence of rosettes, each with five loops of 20 beads. (B) Plot of how the protein–DNA affinity $\varepsilon$ determines the mean search time on a string of rosettes, for different target locations, as investigated in [30]. The target was chosen to be at a location on the DNA which was (i) in the centre of a rosette, (ii) in the loop of a rosette, or (iii) between two of the rosettes. The continuous line shows the search time given according to the scaling theory of eqn (1) for the case of an unstructured protein. Parameters A and $\beta$ and functions $D_1(\varepsilon)$ and $l_1(\varepsilon)$ are taken from fits to simulation data for unstructured DNA as used in Figure 1 in [30]. Times here and in the following plots are in simulation time units; these can be mapped to real units using the Brownian time, defined as $\sigma^2/D_3$ where $\sigma$ and $D_3$ are the diameter and diffusion constant respectively. Using the value of $D_3$ from [2] gives a mapping of approximately 0.64 $\mu$s to one simulation time unit. Note that our simulations are for system volumes and DNA lengths much smaller than a cell, so the search times are much shorter than they would be in reality.

for transcription factors since these bind to the DNA only through a specific binding region (or in some cases through two separate binding regions). Considering periodic boundary conditions (hence avoiding the complications due to confinement), we confirmed that eqn (1) fits our data very well when the DNA is a structureless self-avoiding walk. This fitting requires that the one-dimensional diffusion constant of the protein when bound to the DNA is measured from the simulated trajectories. However, we also showed that DNA conformations play an important role in the dynamics of facilitated diffusion, and, if the DNA has internal structure, e.g. it is looped, then eqn (1) cannot quantitatively capture the mean search time. As we discuss further in the next section, we find that the search time depends strongly on the location of the target in relation to the loops. Also considered in [30] was the effect of sequence heterogeneity, which leads to a rugged, rather than smooth, potential profile for the one-dimensional sliding of the protein.

DNA looping strongly affects the search process

The main contribution of our recent simulations [30] was to highlight that the conformation of the polymer on which proteins need to slide to look for their target may have very large quantitative effects on the search time. The importance of DNA configuration has also been shown experimentally by van der Broek et al. [31], who measured the rate at which a restriction enzyme can cleave a DNA molecule at its target site. By stretching the DNA between two beads, and changing the separation of these beads, they could probe an extended (straight) segment of DNA, a relaxed (coiled) DNA molecule and a DNA molecule which has been squeezed. They found that the rate of cleavage changes depending on both the amount of coiling of the DNA, and the salt concentration (which controls the DNA–protein interaction strength).

In our simulations [30], we focused in particular on the case of a string of rosettes, where by rosette we mean a cloud of loops in the genome (see the sketch in Figure 2A). The rosette structure is suggested by many biological models of both prokaryotic and eukaryotic DNA [32–34]. In such models, DNA loops form due to the presence of molecular ties, which are architectural DNA-binding proteins that have both an affinity for the genome and a tendency to dimerize or multimerize. Although the identity of the ties is still controversial (these might be condensins in bacterial cells, polymerases in transcriptionally active cells or other proteins), their presence is undisputed [35]. Provided that the
Regions of high DNA density can act as traps for searching proteins

(A) Plots of the mean times that a protein spends performing one-dimensional (top) and three-dimensional (middle) diffusion during each search round, and the mean number of search rounds (bottom) required to find a target on a DNA formed into a string of rosettes, as a function of the DNA–protein affinity. Points show results for different target positions relative to the rosettes. The continuous lines show fits to the scaling theory given by eqn (1). (B) Probability distribution of the length of jumps performed by the protein as measured from the simulations. The jump length is defined as the distance along the DNA contour, between the point where the protein leaves and then returns to the DNA when it undergoes an episode of three-dimensional diffusion. Here the DNA–protein affinity is 2.16 $k_BT$, but the same pattern is seen for other values. (C) When the search process is dominated by one-dimensional sliding (high DNA–protein affinity), the rosettes act as traps for the protein. This plot shows the position of the protein on the DNA ($s$ labels the DNA beads) as a function of time for $\epsilon = 5.9 \, k_BT$. The broken lines indicate segments of the DNA belonging to different rosettes. The protein spends a long time in the vicinity of one rosette, before quickly moving to another. If the target is in the centre of a rosette, the protein will find it quickly once it encounters that rosette.

free energy gained via tie dimerization exceeds the loss due to DNA looping, rosettes such as those in Figure 2(A) can form [36].

Figure 2(B) shows the results from [30] for the mean search time as a function of the protein–DNA affinity (which determines the sliding length). Three possible target locations are considered: the centre of a rosette, the middle of a loop in a rosette or the DNA linking two successive rosettes. As the scaling theory in eqn (1) assumes a structureless and translationally invariant DNA, it predicts no dependence on the target location. However, in the loop network of Figure 2(A), this is far from being the case. When the protein–DNA interaction is small, then it takes much longer to find a target when it is in the centre of a rosette. Qualitatively this is due to the crowding of the DNA surrounding the target, which slows down three-dimensional diffusion (the dominant transport mode at low affinity). On the other hand, when the protein–DNA interaction is large, sliding along the genome becomes favoured, and we find that the rosette centre forms a hub which is easy to reach from many different locations on the network, hence the mean search time is reduced when the target is there. In [30], we showed that this intuitive statement can be quantified by looking at the DNA location closest to the protein position: we found that when the interaction with the DNA is large, the protein becomes trapped for a long time in each of the rosettes, and only rarely moves from one rosette to a neighbouring one (Figure 3C). Consequently, for large protein–DNA affinity, the centre of each rosette is visited more often by the protein than other parts of the DNA (results not shown). It is interesting to note that in this case, the theory for a structureless DNA (Figure 2B) predicts a slightly slower search than is the case for a target in the centre of a rosette. Although we used a random initial position for the protein, one might expect that, for a protein starting in the vicinity of the rosette which has the target in its centre, the search time would be reduced further (since in this regime, the rosette acts as a kinetic trap). Indeed, there is some evidence that the gene coding sequence for a transcription factor is often in close proximity to its target [37,38].

Of course, eqn (1) gives a poor estimate of the mean search time for the rosette case, because, as noted above, none of the parameters which feature in it depend on the position of the target. Figure 2(B) shows the best fit which can be obtained when the target is in the middle of a loop on the rosette: a closer look at the data (Figure 3A)
allows one to trace the discrepancy to the poor estimate of the number of search rounds in eqn (1). Ultimately, this is due to the occurrence of a huge number of correlated jumps and intersegmental transfers when the protein is close to the rosette centre; these not being accounted for by the simple scaling theory. In summary, for facilitated diffusion on a string of rosettes, the take-off and landing points in for successive three-dimensional excursions are far from being uncorrelated, and analytic models which do not take this into account cannot correctly predict search times. Evidence for the correlations is shown in the probability distribution function for the distance between the jump take-off and landing points (Figure 3B); this shows marked peaks which appear periodically, with a mutual separation close to the loop size. The increase in probability of jumps which are multiples of the loop length can be explained, for example, by considering the case where the DNA–protein affinity is small. As noted above, in this regime, the loops act as a shield for the interior of the rosette, and the protein is likely to make contact with the DNA at the outermost point of one of the loops; the next contact is likely to also be at the outermost point of one of the loops, and so will be an integer number of loop lengths away along the DNA contour.

Some open questions
We hope that our short survey of the computational studies of facilitated diffusion, the intracellular search of a protein for its target binding site on the genome, has shown that this interesting field is still in its infancy. In the future, larger-scale simulations of the search process have the potential to much improve our understanding of the biophysics of this process, and of intracellular dynamics in general. Specifically, in the present review, we have focused on the fact that the conformation of the genome has a potentially very large effect on the efficiency of facilitated diffusion, and this previously underappreciated fact needs to be explored much more in the future. For instance, does the presence of entangled regions or DNA knots affect the mean search time in a similar way to the presence of loops?

More generally, within a living cell, there are a vast number of proteins interacting with the DNA, trying to find a protein-specific target (e.g. the case of transcription factors), acting to fold the genome (e.g. the case of architectural proteins) or performing a biological function (e.g. polymerases and other nanomachines). These ensembles of proteins could cooperate to find a target more rapidly, or could create traffic jams on the DNA that are likely to hinder the efficiency of facilitated diffusion [5]. These effects are outside the scope of current simulations.

The interior of a living cell is qualitatively different from the buffer of an in vitro experiment, if nothing else because of the much enhanced macromolecular crowding found in the cell, which will affect the dynamics of three-dimensional excursions in facilitated diffusion. Just like protein traffic and co-operative search, the effect of crowding is difficult to understand without large-scale simulations, and we expect that numerical study will teach us a lot about the quantitative effect on the search process of these and similar phenomena.

Funding
We acknowledge the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/I034661/1] for funding. M.E.C. is funded by the Royal Society.

References

©The Authors Journal compilation ©2013 Biochemical Society

Received 10 September 2012
doi:10.1042/BST20120234