A Markov model for the dynamics of the nucleosome

B van Opheusden[1]**, F Redig**[2] **and H Schiessel**[1]

[1] Lorentz Instituut, Universiteit Leiden, Leiden, The Netherlands
[2] Delft Institute of applied mathematics, Mekelweg 4, 2628 CD Delft, The Netherlands

E-mail: opheusden@lorentz.leidenuniv.nl

Abstract

We study the dynamics of the nucleosome, the most abundant DNA protein complex in eukaryotic cells. The dynamics consist of two mechanisms, site exposure and sliding, which are crucial for making DNA accessible to DNA-binding proteins. The intertwining of both effects leads to a rich stochastic process that has not been studied before. Within the assumptions of our model, nucleosomes perform a symmetric random walk along the DNA strand. We investigate how the diffusion constant depends on the relative rates of site exposure and sliding.

PACS numbers: 02.50.Ga, 87.15.kj, 87.15.H−

(Some figures may appear in colour only in the online journal)

1. Introduction

The DNA of eukaryotic cells (cells of animals, plants and fungi) has typically macroscopic lengths but needs to fit inside micron-sized cell nuclei. To achieve this and to have the genetic information accessible, eukaryotic DNA is hierarchically folded with the help of proteins [1]. On the first level, DNA is wrapped around millions of protein cylinders resulting in a string of DNA spools, called nucleosomes. The core of each spool is a cylinder composed of eight histone proteins and is wrapped by a 147 base pairs (bp) long stretch of DNA. A short stretch of unbound DNA, the linker DNA, connects to the next protein spool. It is known from the nucleosome crystal structure [2] that the DNA is bound to the protein core at 14 binding sites at which the minor groove of the DNA double helix faces the cylinder. This defines the binding path, a left-handed superhelix of one and three quarter turns. The details of higher order structures beyond the nucleosome are still a matter of debate [3] and are not discussed here any further.

With three quarters of eukaryotic DNA being wrapped into nucleosomes, the question arises how other DNA-binding proteins can bind to their target DNA sequence if it happens to be located inside a wrapped DNA portion. As the result of the proximity of the impenetrable surface of the protein cylinder, steric exclusion makes such target sequences inaccessible. However, spontaneous fluctuations of the nucleosomes allow transient access. This is possible

via two dynamic modes: the site exposure mechanism [4–12], where the DNA partially unwraps from the protein core, and nucleosome sliding [13–18] where the protein cylinder moves as a whole along the DNA.

The site exposure mechanism is based on thermal fluctuations where nucleosomal DNA unwraps spontaneously from either end of the wrapped DNA portion as the result of the sequential opening of binding sites. For each unbinding event, the DNA has to pay the price to open a binding site but reduces its bending energy since the corresponding 10 bp long stretch can straighten. The change in the net energy is very small, on the order of the thermal energy [19]. Experimental methods detect the site exposure mechanism either through measuring protein accessibility to nucleosomes [4–8] or, more recently, through fluorescence energy transfer between dyes placed on the nucleosomal DNA and/or the protein cylinder [9–12].

Nucleosome sliding along DNA is caused by small defects that spontaneously form in the wrapped DNA. There is experimental evidence [17, 18] that those defects are small 1 bp twist defects where an extra or a missing base pair is located between two neighbouring binding sites [20, 22, 23]. The corresponding piece of DNA has to be over- or understretched and over- or undertwisted to accommodate the defect. Twist defects are produced at either end of the wrapped portion and then start to diffuse through the wrapped DNA portion. Since there are 14 binding sites between the DNA and the protein core and since a twist defect is localized between two binding sites, the defect has 13 possible positions on the nucleosome. If the defect eventually falls off at the same end where it had been produced, then the nucleosome does not change its position on the DNA. If, on the other hand, it diffuses through the whole wrapped portion and falls off at the other end, then the nucleosome makes a step by 1 bp along the DNA. The cost of a twist defect has been estimated to be about $9\,k_BT$, where T denotes the absolute temperature and k_B the Boltzmann constant [20]; this means that the probability of having more than one defect on a nucleosome is negligible.

So far the two different dynamical modes of the nucleosome have been treated separately in the literature, i.e. the site exposure mechanism has been considered in the absence of defects [19] and nucleosome sliding in the absence of site exposure [20]. The mathematics behind these two separate treatments is fairly straightforward. In reality, however, the two mechanisms are intertwined. For instance, a defect inside the nucleosome might disappear without moving to another binding site when DNA unwraps beyond the location of that defect. There have been studies [21] which investigate the influence of site exposure on the sliding mechanism, but our purpose is to combine both into one dynamical model.

2. Model description and approximations

2.1. The Markov model

We consider a single nucleosome attached to an infinitely long DNA strand, from the moment a defect originates at either end of nucleosome, until the defect leaves the structure. We want to answer the following question: *given that a defect enters at one end, what is the probability for it to exit at the other end?*

We model the system as a stationary continuous time Markov chain, with a state space determined by the following three parameters (see also figure 1) for $t \in [0, \infty)$:

- The number of unwrapped binding sites on the left end, labelled a_t,
- The number of unwrapped binding sites on the right end, labelled b_t,
- The 'position' of the defect, labelled D_t. Formally, D_t is defined as the number of binding sites from the defect up to and including the leftmost binding site.
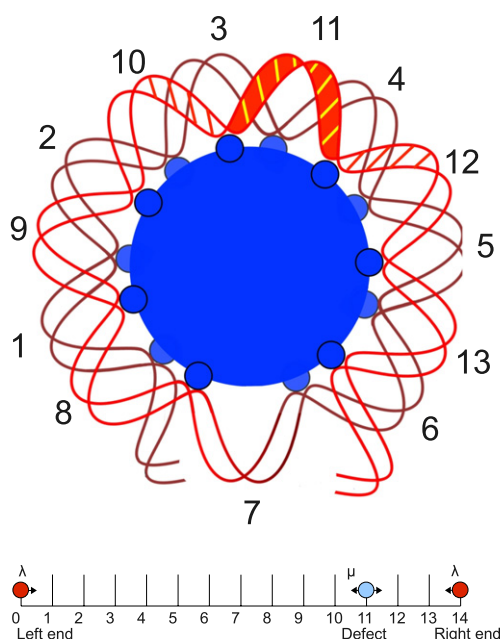
Figure 1. Visualisation of our model for nucleosome sliding and site exposure. The nucleosome (blue) is fully wrapped with DNA (red), and there is a twist defect, represented by the yellow stripes. This state is represented by three integers; the position of the defect and the number of unwrapped loops from either side. In this case, $a_t = 0$, $b_t = 0$, and $D_t = 11$.

These are not free parameters, because a_t and b_t have to be non-negative integers and $a_t \leqslant D_t \leqslant 14 - b_t$ for all time. The Markov chain starts with $a_0 = b_0 = 0$, $D_0 = 1$, and ends whenever $D_t = a_t$ or $D_t = b_t$.

Let us now consider the dynamics of the Markov chain. We assume that site exposure and sliding evolve independently, so we consider their transitions separately. Unwrapping or wrapping at one end causes a_t or b_t to increase or decrease by 1, so there are allowed transitions $a_t \to a_t \pm 1$ and $b_t \to b_t \pm 1$. The sliding process allows the defect to jump to a neighbouring DNA segment, so we also have $D_t \to D_t \pm 1$.

The rates for these transitions depend on a_t, b_t and D_t as well as the underlying sequence of the DNA strand [4, 5, 13, 14], but incorporating this into the model would make it too complicated. We want to examine the effect of site exposure versus sliding, so we set all wrapping and unwrapping rates to a constant λ, and all sliding rates to another constant μ. Changes in environmental conditions that favour site exposure over sliding or vice versa can be modelled by varying the ratio between λ and μ. Since sliding is always much faster than site exposure, we will allow approximations that are valid for $\lambda \ll \mu$.

With this approximation the model is symmetric, which implies that the nucleosome performs a symmetric random walk along the DNA. The rate at which the nucleosome moves is equal to the rate A at which defects are generated times the probability P_{success} that such a defect leaves the structure at the other end than it entered. Since such a move results in a 1bp step, the diffusion constant is equal to

$$D = \tfrac{1}{2} A (1\text{bp})^2 P_{\text{success}}. \tag{1}$$
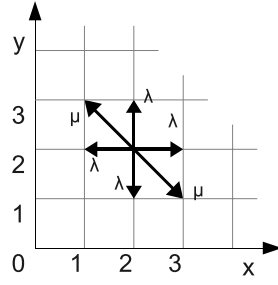
Figure 2. The main approximation we use in this paper ignores the constraints on *a*, *b* and *D* due to the finite size of the nucleosome and formulates the problem in terms of relative distances *x*, *y*. This stochastic process is a continuous time random walk in the quarter plane with uniform transition rates. The process starts with $x = 1$, $y = 13$, and it results in a successful move when it reaches the *x*-axis without hitting the *y*-axis.

Both *A* and P_{success} depend on environmental conditions, but in this study we will focus on the latter term. This probability also depends on the starting condition. The defect will always be generated at the first or last DNA segment, but there could be binding sites that are unwrapped. We derive formulas that are valid for any initial state, with emphasis on the fully wrapped state.

2.2. Infinite lattice approximation

This process turns out to be too complicated to calculate in full detail, so we have to make some approximation. We remove the restriction that a_t and b_t have to be non-negative, but we still start with $a_0 = b_0 = 0$. Physically, this means the DNA can wrap arbitrarily many times around the nucleosome. Since those nonphysical moves involve wrapping and unwrapping events, this approximation is expected to behave best in the regime of small λ.

Given the infinite lattice approximation, the process has become translation-invariant. Therefore, the whole dynamics can be described in terms of the relative distances

$$x_t := D_t - a_t, \, y_t := 14 - b_t - D_t. \tag{2}$$

The process starts with $(x_0, y_0) = (1, 13)$, and it stops whenever $x_t = 0$ or $y_t = 0$. The allowed transitions for (x_t, y_t) are (see also figure 2)

$$
\begin{aligned}
(x, y) &\rightarrow (x \pm 1, y) & \lambda \\
(x, y) &\rightarrow (x, y \pm 1) & \lambda \\
(x, y) &\rightarrow (x \pm 1, y \mp 1) & \mu.
\end{aligned} \tag{3}
$$

In other words, the transition rates

$$p_{ij} = \lim_{\tau \downarrow 0} \frac{\mathbb{P}[(x_{t+\tau}, y_{t+\tau}) = (x + i, y + j) | x_t = x, y_t = y]}{\tau} \tag{4}$$

are given by $p_{0,1} = p_{0,-1} = p_{1,0} = p_{-1,0} = \lambda$ and $p_{1,-1} = p_{-1,1} = \mu$. For convenience, we normalize time such that

$$\sum_{ij} p_{ij} = 1, \tag{5}$$

so that p_{ij} is the probability for the random walk to move in the direction (i, j) given that it makes a move. The dynamics are most clearly described by the generator

$$
\begin{aligned}
Lg(x, y) &= \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}[g(x_t, y_t) | x_0 = x, y_0 = y]|_{t=0} \\
&= \sum_{ij} p_{ij} g(x + i, y + j) - g(x, y),
\end{aligned} \tag{6}
$$

where $\mathbb{E}$ denotes the expectation value under the process. The main object of study is the function $f(x, y)$, the probability that the random walk started at (x, y) hits the x-axis before it hits the y-axis. In particular, the probability of success is given by

$$P_{\text{success}} = f(1, 13). \tag{7}$$

Because the random walk has no drift, the hitting time of either axis is finite with probability 1, so we can write

$$f(x, y) := \mathbb{P}(y_\tau = 0 | x_0 = x, y_0 = y), \tag{8}$$

with τ the hitting time

$$\tau := \inf\{t \geqslant 0 : x_t = 0 \vee y_t = 0\}. \tag{9}$$

It is easy to see that $f(x, y)$ is increasing in x and decreasing in y. By the first step analysis, we obtain that f is the solution to the Dirichlet problem

$$\forall y > 0 : f(0, y) = 0,$$
$$\forall x \geqslant 0 : f(x, 0) = 1,$$
$$\forall x, y > 0 : Lf(x, y) = 0, \text{ i.e.}$$
$$f(x, y) = \sum_{i,j} p_{ij} f(x + i, y + j). \tag{10}$$

Moreover, the inherent symmetry in the transition probabilities immediately implies that

$$\forall x, y > 0 : f(x, y) + f(y, x) = 1. \tag{11}$$

3. Discrete results

3.1. Limiting cases: simple symmetric random walk

The Dirichlet problem is easy to solve in the absence of site exposure, which means $\mu = \frac{1}{2}$ and $\lambda = 0$. The motion is restricted to the diagonal $\{x + y = 14\}$, and the one-dimensional Dirichlet problem is trivial. In general, $f(x, y) = \frac{x}{x+y}$, so the probability of success equals $f(1, 13) = \frac{1}{14}$. The simplest non-degenerate random walk is the simple symmetric random walk, for which $\mu = 0$ and $\lambda = \frac{1}{2}$. This corresponds to a nucleosome for which no sliding occurs, only site exposure. In this case, the solution can also be computed exactly, because the eigenvalues and eigenfunctions of the generator are known. This information enables us to solve the Dirichlet problem using a theorem by Chung [24]. For the sake of completeness, in the appendix, we prove this theorem for general reversible random walks on finite graphs. Here, we are working on an infinite state space, but we can reduce to the finite problem by means of a limiting procedure.

We restrict the process to a finite box, stopping the process when either x or y becomes equal to some fixed N. We then calculate the probability $f_N(x, y)$ that the random walk leaves this box through the x-axis. As $N \to \infty$, it becomes less likely that the random walks exits through the far end of the box, so f_N converges pointwise to f. In other words, to find $f(x, y)$, we define

$$\tau_N := \inf\{t \geqslant 0 : x_t = 0 \vee y_t = 0 \vee x_t = N \vee y_t = N\}, \tag{12}$$

compute

$$f_N(x, y) := \mathbb{P}(y_{\tau_N} = 0 | x_0 = x, y_0 = y) \tag{13}$$

and let $N \to \infty$. Now we have to solve the Dirichlet problem with additional boundary conditions $f(x, N) = f(N, y) = 0$. Because the simple symmetric random walk is homogeneous inside the box, we can put Chung's theorem in an easier form.

Theorem 1. *The solution to the Dirichlet problem on a finite subset $S \subset \mathbb{N}^2$ with boundary condition $g : \delta S \to \mathbb{R}$ is given by*

$$f(x, y) = \frac{1}{4} \sum_{i \in \mathcal{I}} \frac{1}{\lambda_i} \sum_{S \ni (x', y') \sim (x'', y'') \in \delta S} \phi_i(x, y) \phi_i(x', y') g(x'', y'') \tag{14}$$

where $\{(\phi_i, \lambda_i), i \in \mathcal{I}\}$ is an orthonormal eigensystem of the generator on S.

We will choose S to be the $N \times N$ box, and $g(x, 0) = 1$ and zero otherwise. The generator of the simple symmetric random walk is the discrete Laplacian given by

$$\Delta g(x, y) := \frac{g(x, y + 1) + g(x, y - 1) + g(x + 1, y) + g(x - 1, y)}{4} - g(x, y). \tag{15}$$

The orthonormal eigensystem of Δ is easily computed to be

$$\phi_{mn}(x, y) = \frac{2}{N} \sin\left(\frac{\pi m x}{N}\right) \sin\left(\frac{\pi n y}{N}\right),$$
$$\lambda_{mn} = 1 - \frac{1}{2} \cos\left(\frac{n\pi}{N}\right) - \frac{1}{2} \cos\left(\frac{m\pi}{N}\right), \tag{16}$$
$$m, n \in \{1, 2, \ldots, N - 1\}.$$

We can insert this into the statement of the theorem to obtain

$$f_N(x, y) = \frac{1}{N^2} \sum_{\substack{m=1 \\ m \text{ odd}}}^{N-1} \sum_{n=1}^{N-1} \frac{\sin\left(\frac{\pi m}{N}\right) \sin\left(\frac{\pi n}{N}\right) \sin\left(\frac{\pi m x}{N}\right) \sin\left(\frac{\pi n y}{N}\right)}{\left(1 - \cos\left(\frac{\pi m}{N}\right)\right)\left(1 - \cos\left(\frac{n\pi}{N}\right)/2 - \cos\left(\frac{m\pi}{N}\right)/2\right)}. \tag{17}$$

As $N \to \infty$, the summations become integrals. The first summation is only over the odd integers, which gives an overall factor of 1/2. Also, because the summand involves functions of $\pi n/N$ rather than n/N, there is a factor of $1/\pi^2$. Together, we obatin

$$f(x, y) = \frac{1}{2\pi^2} \int_0^\pi \int_0^\pi \frac{\sin(u) \sin(v) \sin(xu) \sin(yv)}{(1 - \cos(u))(1 - \cos(u)/2 - \cos(v)/2)} \, du \, dv. \tag{18}$$

In the appendix, we show how to solve this integral algebraically. The physically important number is the value of f for $x = 1, y = 13$, and this turns out to be

$$f(1, 13) = 42\,344\,121 - \frac{1\,198\,449\,065\,536}{9009\pi} \approx 0.048\,969\ldots \tag{19}$$

3.2. Asymptotic relations

The method employed in the last section does not generalize from the simple symmetric random walk to arbitrary λ and μ. In order to use Chung's theorem, one needs to know the eigenfunctions of the normalized Laplacian. Unfortunately, for a general weighted graph, determining the eigenfunctions is just as hard as solving the Dirichlet problem directly.

It is possible to find $f(x, y)$ exactly, using an elaborate technique developed by Malyshev [25]. However, this method is computationally difficult, and it only yields an integral expression. Alternatively, one can use ideas from harmonic function theory [26], but that is also quite involved. Therefore, we take a different route. First, we determine the asymptotic behaviour of $f(x, y)$ for $y \gg x$, which will turn out to be $f(x, y) \approx cx/y$, where c depends on μ and λ. Then, we switch to a continuum version of the process, which does not change the asymptotic behaviour, and allows us to compute $c(\mu, \lambda)$ analytically. We can approximate $f(1, 13)$ either by $\frac{c}{14}$ or by the solution for $f(1, 13)$ in the continuum process. Neither will be an exact match, but the dependence on μ and λ is qualitatively the same for each.

Let us now ask what happens to $f(x, y)$ as $y \to \infty$. Because $f(x, y)$ is decreasing in y, the limit $L_x := \lim_{y \to \infty} f(x, y)$ exists for all $x \geqslant 0$. From the Dirichlet condition (10), it follows that

$$L_x = \frac{L_{x-1} + L_{x+1}}{2}, \tag{20}$$

which means that $x \to L_x$ is linear. Since L_x is also bounded, and $L_0 = 0$ by definition, L_x has to vanish for all $x \geqslant 0$. For later purposes, it is convenient to rewrite the limit $\lim_{y \to \infty} f(x, y)$ in terms of generating functions, because for the generating functions there is a useful functional equation. So we define the functions $V : \mathbb{R}^2 \to \mathbb{R}$, and $V_{1,2} : \mathbb{R} \to \mathbb{R}$ by

$$V(u, v) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(m, n) u^m v^n, \tag{21a}$$

$$V_1(u) = \sum_{m=1}^{\infty} f(m, 1) u^m, \tag{21b}$$

$$V_2(v) = \sum_{n=1}^{\infty} f(1, n) v^n. \tag{21c}$$

Because the coefficients of $V(u, v)$ are probabilities, they are bounded, and as a result $V(u, v)$ converges absolutely and uniformly on compact subsets of $\mathcal{D} := \{(u, v) \in \mathbb{R}^2 : |u| < 1, |v| < 1\}$. The same applies to $V_1(u)$ and $V_2(v)$. For $u, v \in \mathcal{D}$, we can use the Dirichlet condition (10) to write

$$
\begin{aligned}
V(u, v) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(m, n) u^m v^n \\
&= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \left(\sum_{ij} p_{ij} f(m+i, n+j) \right) u^m v^n. \\
&= \sum_{ij} p_{ij} u^{-i} v^{-j} \left(\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} f(m+i, n+j) u^{m+i} v^{n+j} \right). \tag{22}
\end{aligned}
$$

The term inside the brackets is equal to $V(u, v)$ plus a couple of boundary terms. These boundary terms can be simplified, noting that

$$V_1(u) + V_2(u) = \sum_{n=1}^{\infty} [f(n, 1) + f(1, n)] u^n = \sum_{n=1}^{\infty} u^n = \frac{u}{1 - u}. \tag{23}$$

Finally, after some algebra we arrive at the functional equation

$$D(u, v)V(u, v) + uv[(\mu v + \lambda)V_2(v) - (\mu u + \lambda)V_2(u)] = \frac{u^2 v}{1 - u}(\mu v - \mu u + \lambda v - \lambda), \tag{24}$$

where

$$D(u, v) = uv - \lambda(u^2 v + uv^2 + u + v) - \mu(u^2 + v^2). \tag{25}$$

The limit L_1 can be rewritten using Abel's theorem, since

$$\lim_{y \to \infty} f(1, y) = \lim_{u \to 1} (1 - u) V_2(u). \tag{26}$$

To calculate this last limit, we multiply the functional equation by $1 - u$, and let $u \to 1$. Then, we obtain

$$(v - 1)^2 \lim_{u \to 1} (u - 1) V(u, v) + v \lim_{u \to 1} (u - 1) V_2(u) = v(v - 1). \tag{27}$$

Because $f(m, n) \in [0, 1]$ for all m and n, we can estimate $V(u, v) \leqslant |uv/((1 - u)(1 - v))|$. This means the first term in this equation will go to zero as $v \to 1$. The third term also vanishes, so $\lim_{u \to 1}(u - 1)V_2(u) = 0$. Therefore, we have arrived at the simple conclusion

$$\forall x \geqslant 0 : \lim_{y \to \infty} f(x, y) = 0. \tag{28}$$

Now that we know that $f(x, y)$ decays to zero as $y \to \infty$, it is natural to ask how fast this decay is. If $\mu = \frac{1}{2}$, then $f(x, y) = x/(x + y) \approx \frac{x}{y}$. In general, it holds that $f(x, y) \approx \frac{cx}{y}$, where c depends on the parameters μ and λ. Formally, we have to be careful, since the limit $\lim_{y \to \infty} y f(x, y)$ does not necessarily exist. There could be oscillations in $f(x, y)$ of magnitude proportional to $1/y$, which get amplified by the prefactor y. However, since such oscillations do not occur in numerical simulations, we choose to ignore that problem and try to find the solution under the assumption that the limit exists.

As before, we only have to consider the case $x = 1$, because the rest follows from the recurrence relation and the fact that $f(x, y)$ vanishes as $y \to \infty$. Again, we use Abel's theorem to write

$$\lim_{y \to \infty} y f(1, y) = \lim_{u \to 1}(1 - u)\frac{dV_2(u)}{du}. \tag{29}$$

To obtain an expression for the latter limit, we differentiate the functional equation with respect to v, substitute $v = u$ and multiply with $1 - u$. This gives

$$(1 - u)\frac{\partial}{\partial v}[D(u, v)V(u, v)]|_{v=u} + \mu u^2(1 - u)V_2(u) + (\mu u^3 + \lambda u^2)(1 - u)\frac{dV_2(u)}{du}$$
$$= (2\lambda + \mu)u^3 - \lambda u^2. \tag{30}$$

In the limit $u \to 1$, the second term on the left-hand side vanishes. Using the product rule on the first term, and noting that $V(u, u) = \frac{u^2}{2(1-u)^2}$ by symmetry, we obtain

$$\lim_{u \to 1}\left[\frac{2\lambda}{\mu + \lambda}\left(\frac{1}{2} - (1 - u)^3\left.\frac{\partial V(u, v)}{\partial v}\right|_{v=u}\right) + (1 - u)\frac{dV_2}{du}\right] = 1. \tag{31}$$

Since $\frac{1}{2} \leqslant f(m, n) \leqslant 1$ as long as $m \geqslant n$, we can estimate the derivative of $V(u, v)$ from above and below:

$$\frac{1}{8} \leqslant \lim_{u \to 1}(1 - u)^3\left.\frac{\partial V(u, v)}{\partial v}\right|_{v=u} \leqslant 1. \tag{32}$$

Combining these estimates with the equation above, we see that

$$\frac{1 - \frac{7}{2}\lambda}{1 - 2\lambda} \leqslant \lim_{u \to 1}(1 - u)\frac{dV_2(u)}{du} \leqslant \frac{1}{1 - 2\lambda} \tag{33}$$

Since $\lambda \leqslant \frac{1}{4}$, the lower bound is always positive, and there exists a constant c such that

$$\lim_{y \to \infty} y f(x, y) = cx \tag{34}$$

for all $x \geqslant 0$.

On a side note, the asymptotic relation $f(x, y) \approx \frac{cx}{y}$ as $y \to \infty$ is quite general amongst random walks in the quarter plane. Using the same method as above, we have been able to show it for nearest-neighbour random walks which have symmetric transition rates, no drift, and for which $p_{1,1} + p_{-1,-1} < \frac{1}{5}$. We believe the assumptions of nearest-neighbour jumps and symmetry can be relaxed, but the no-drift condition is crucial. The gradient of $D(u, v)$ at the point $(1, 1)$ is equal to the drift, so only in the case of no drift does $D(u, v)$ have an isolated zero there. An example with a negative drift is investigated by Godrèche [27].

4. Continuum limit

The asymptotic analysis in the previous section provides us with an upper and a lower bound on the function $c(\mu, \lambda)$, but it does not indicate how to find its exact value. This is not a result of insufficient analysis, but a typical feature of these kinds of problems (for example, see [27]). Fortunately, there is also a typical trick to solve this issue. It can be shown that the asymptotic behaviour of $f(x, y)$ will not change if we switch to the continuum limit of the same random walk. The discrete Dirichlet problem then turns into a differential equation, which is analytically solvable in our case.

In the continuum limit, the graph of the random walk is modified by adding extra vertices and edges, in such a way that the new graph is still isomorphic to the original one. Additionally, the transition rates are altered to ensure that the new random walk evolves at the same time scale as before. Iterating this procedure results in a finer and finer grid, and in the limit we obtain a stochastic process on $\mathbb{R}^2$, which turns out to be a stretched version of Brownian motion.

To make this notion precise, we view the process $X_t = (x_t, y_t)$ as a process on $\mathbb{R}^2$, and set $X_t^{(N)} := \frac{1}{N} X_{N^2 t}$. As $N \to \infty$, these processes converge in some sense to a continuous process X_t^c on $\mathbb{R}^2$. By the Trotter–Kurtz theorem [28], it suffices to show that the generators converge on a large enough set of functions. The generator of $X_t^{(N)}$ is

$$
L_N g(x, y) = N^2 \left[\sum_{ij} p_{ij} g \left(x + \frac{i}{N}, y + \frac{j}{N} \right) - g(x, y) \right]. \tag{35}
$$

As $N \to \infty$, these generators converge to

$$
L_c = \frac{1}{2} \sum_{i,j} i(i - j) p_{ij} \left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y} \right)^2 + \frac{1}{2} \sum_{i,j} i(i + j) p_{ij} \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right)^2, \tag{36}
$$

for sufficiently regular functions. It turns out that if we restrict to the set of infinitely differentiable functions with compact support and uniformly bounded derivatives, then all the requirements of the Trotter–Kurtz theorem are satisfied.

Because the event $y_\tau = 0$ is invariant under spatial and temporal scaling, we can interpret the continuum limit in a different way. If we define the stopping times τ_N and τ_c as well as the hitting probabilities $f_N(x, y)$ and $f_c(x, y)$ analogous to τ and f, then we can write

$$
\begin{aligned}
f_N(x, y) &= \mathbb{P}(y_{\tau_N}^{(N)} = 0 | x_0^{(N)} = x, y_0^{(N)} = y) \\
&= \mathbb{P}(y_\tau^{(N)} = 0 | x_0^{(N)} = x, y_0^{(N)} = y) \\
&= \mathbb{P}(y_\tau = 0 | x_0^{(N)} = x, y_0^{(N)} = y) \\
&= \mathbb{P}(y_\tau = 0 | x_0 = Nx, y_0 = Ny) \\
&= f(Nx, Ny).
\end{aligned} \tag{37}
$$

Since the processes $X_t^{(N)}$ converge to X_t^c, f_N converges to f_c at least pointwise, and in the limit $N \to \infty$ we have

$$
f_c(x, y) = \lim_{N \to \infty} f(\lfloor Nx \rfloor, \lfloor Ny \rfloor) \tag{38}
$$

for all $x, y \in \mathbb{R}$ ($\lfloor \cdot \rfloor$ denotes the integer part of a real number). This convergence allows us to compute c, since

$$
\begin{aligned}
\lim_{n \to \infty} n f_c(1, n) &= \lim_{n \to \infty} \lim_{N \to \infty} n f(N, nN) \\
&= \lim_{N \to \infty} \lim_{n \to \infty} n f(N, nN) = c, \tag{39}
\end{aligned}
$$

where we have tacitly assumed that the exchange of limits can be justified by a uniform convergence argument. So to determine the function c we can work with the continuous process, for which the hitting time problem is much simpler. Indeed, it is well known [29, 30] that f_c is the solution of the Dirichlet problem

$$
\begin{aligned}
f_c(x, 0) &= 1 \\
f_c(0, y) &= 0 \\
L_c f_c &= 0,
\end{aligned}
\tag{40}
$$

which is the continuous analogue of (10). To solve this differential equation, we define new coordinates

$$
\tilde{x} := \frac{x + y}{\left(\sum_{i,j} i(i - j) p_{ij} \right)^{1/2}}, \qquad \tilde{y} := \frac{x - y}{\left(\sum_{i,j} i(i + j) p_{ij} \right)^{1/2}},
\tag{41}
$$

so that the generator assumes the canonical form of Brownian motion

$$
L_c = \frac{1}{2} \left(\frac{\partial^2}{\partial \tilde{x}^2} + \frac{\partial^2}{\partial \tilde{y}^2} \right).
\tag{42}
$$

This means f_c is a harmonic function of $\tilde{x}$ and $\tilde{y}$, which we will try to construct as the imaginary part of a holomorphic function of $z = \tilde{x} + i\tilde{y}$. The domain of this function is the wedge

$$
\{ z \in \mathbb{C} \setminus \{0\} : -\arctan(\alpha) < \arg(z) < \arctan(\alpha) \},
\tag{43}
$$

where we have defined

$$
\alpha := \left(\frac{\sum_{i,j} i(i + j) p_{ij}}{\sum_{i,j} i(i - j) p_{ij}} \right)^{1/2}.
\tag{44}
$$

This domain can be biholomorphically mapped onto the upper half-plane $\mathcal{H}$ by the function $z \to iz^{(\pi/(2\arctan\alpha))}$. The unique holomorhic function on $\mathcal{H}$ satisfying the correct boundary conditions is $(1/\pi) \ln(z)$, which implies

$$
f_c(x, y) = \frac{1}{2} + \frac{\arctan\left(\alpha \frac{x - y}{x + y} \right)}{2 \arctan(\alpha)}.
\tag{45}
$$

The function c can now be computed by a Taylor expansion of f_c, which gives

$$
c = \frac{\alpha}{(1 + \alpha^2) \arctan \alpha}.
\tag{46}
$$

Finally, because $\alpha = \left(\frac{\lambda}{\lambda + 2\mu} \right)^{1/2} = \mathrm{O}(\sqrt{\lambda})$, in the limit of slow site exposure, $c = 1 - \frac{2\lambda}{3} + \mathrm{O}(\lambda^2)$.

4.1. Comparison with simulation

Using the continuum limit, we managed to compute $c(\mu, \lambda)$ exactly, and we obtained an approximation $f_c(x, y)$. This approximation cannot be exact, since in the special case of simple symmetric random walk we have

$$
f_c(1, 13) = \frac{1}{2} - \frac{2}{\pi} \arctan \left(\frac{12}{14} \right) \approx 0.048\,875,
\tag{47}
$$

whereas earlier we have computed

$$
f(1, 13) = 42\,344\,121 - \frac{1\,198\,449\,065\,536}{9009\pi} \approx 0.048\,969.
\tag{48}
$$

Still, the difference is very small. In figure 3, we compare $f_c(1, 13)$ to a numerical simulation of $f(1, 13)$ and the exact value of $\frac{c}{14}$. The difference between the continuum and discrete
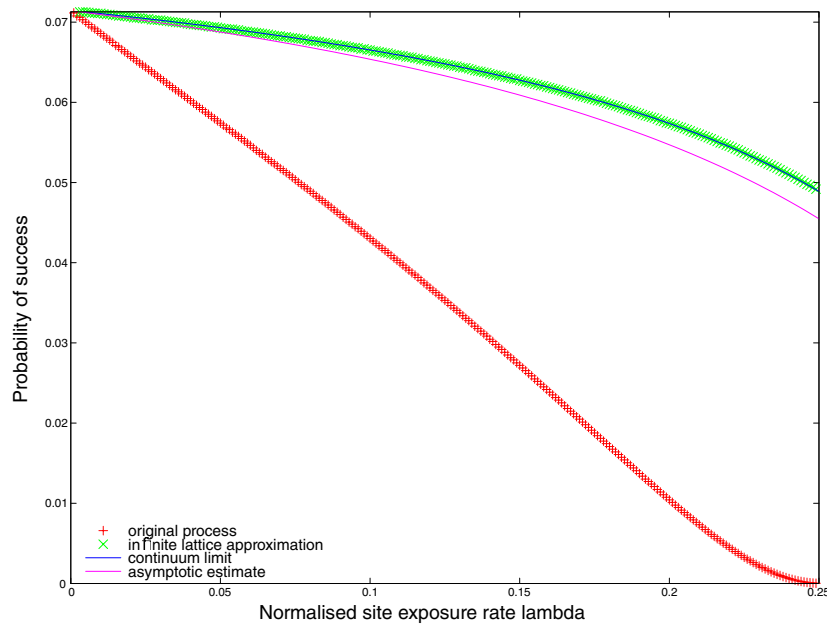
Figure 3. Various approximations for the probability of success are suggested. Shown are numerical results for the infinite lattice approximation and two analytical ways to compute it, one based on the asymptotic analysis and one based on the continuum limit. Neither is exact, but the continuum limit is extremely close. However, the infinite lattice approximation itself is weak, even in the limit of slow site exposure.

process remains negligible for all λ, whereas the asymptotic estimate gets worse. Even near $\lambda = 0$, the slope of the continuum and discrete process match up, but the asymptotic estimate does not.

Unfortunately, the infinite lattice approximation breaks down for all non-zero values of λ. This is caused by the fact that in the infinite lattice approximation, the nucleosome can always wrap and unwrap from both sides. Physically, this makes no sense, because wrapping can only take place if there is space available on the nucleosome. Since this is a site exposure effect, the difference will scale as $O(\lambda)$ for small λ.

4.2. A new model: triangle approximation

Let us try to refine the model to obtain better agreement with the simulation. We consider the same random walk, but now conditioned on the event that $x_t + y_t \leqslant 14$ for all $t \geqslant 0$. This still allows for impossible movements, though. If the nucleosome unwraps at one end, it is possible in this model to wrap at the other end. For this effect to occur two unwrapping/wrapping steps are needed, so we expect its contribution to be $O(\lambda^2)$.

Because the state space of the random walk is now finite, the Dirichlet problem is easy to solve. The generator becomes a matrix, and the Dirichlet problem amounts to a simple exercise in linear algebra. However, this matrix is still 91×91, so the solution will be a complete mess. This approach will yield the correct answer, but it does not provide with any insight on its behaviour. We cannot switch to a continuum limit either, because the event $x_t + y_t \leqslant 14$ is not scale invariant, so the processes $X_t^{(N)} = \frac{1}{N} X_{N^2 t}$ will not converge. Therefore, we have to resort to a numerical simulation again, the results of which are shown in figure 4.
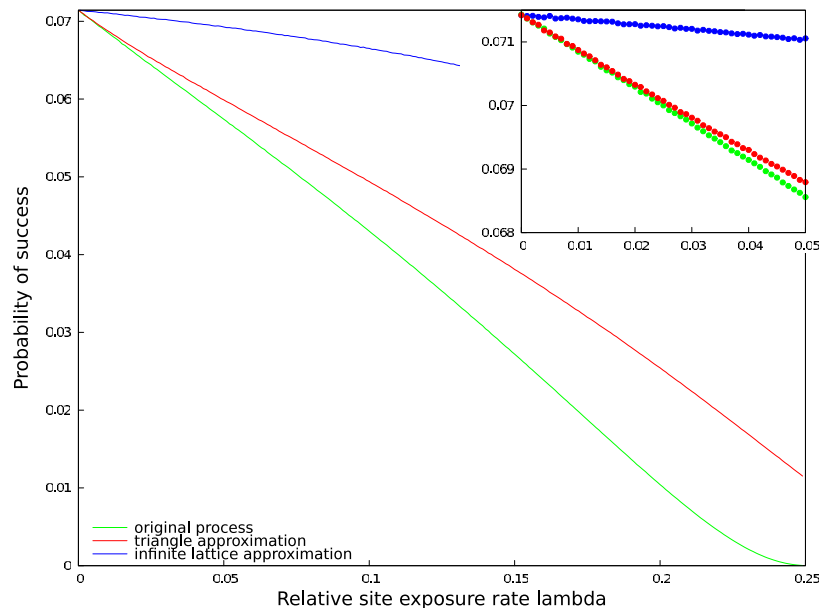
Figure 4. Motivated by the failure of the infinite lattice approximation, we propose a refinement of the model, where the relative positions are restricted to the triangle $\{x+y \leqslant 14\}$. The probability of success can no longer be calculated analytically, so here we show results of a numerical simulation. The triangle approximation is still not exact, especially at large λ, but the slope at $\lambda = 0$ is correct, evident from the inset.

5. Conclusion

We found approximate results for the diffusion constant of nucleosomes along a DNA strand, incorporating both sliding and site exposure. The site exposure effect will always reduce the probability for defects to succeed in moving the nucleosome.

Most of the results were derived using the infinite lattice approximation, which assumes the nucleosome to be infinitely large. In this approximation, we have derived a near-exact expression, which becomes exact in the limit of slow wrapping/unwrapping.

To obtain a mathematically solvable model, we neglected the dependence of the site exposure and sliding rates on the position of the defect and the number of unwrapped loops. These rates also depend quite a lot on the underlying sequence of the DNA. As an extension of the model, it would be interesting to see how position-dependent rates will affect the diffusion constant. We could also investigate the effect of forces on the nucleosome. That would introduce an asymmetry in the problem, so the nucleosome will perform a random walk with drift. Moreover, due to the force, unwrapping will be favoured over wrapping, and the nucleosome will be removed from the DNA strand after a finite time. Results on force-induced nucleosome movement and desorption have been shown by Chou [21].

Appendix A. Dirichlet problems for finite Markov chains

In this appendix, we consider a Markov chain with finite state space X. Moreover, we will assume that it is reversible with rates $r_{x \to y}$ and stationary measure $\pi(x)$. The generator of such

a process has the simple form

$$Lg(x) := \sum_{y \in X} r_{x \to y}[g(y) - g(x)]. \tag{A.1}$$

The Markov chain has an associated weighted graph defined by $x \sim y \iff r_{x \to y} \neq 0$ and edge weights $w_{xy} := \pi(x)r_{x \to y}$. We can formulate the Dirichlet problem in terms of this graph. Let $S \subset X$ be a connected subgraph, and $\delta S = \{x \notin S : \exists y \in S : x \sim y\}$. The Dirichlet problem amounts to finding, for a given $g : \delta S \to \mathbb{R}$, a function $f : S \cup \delta S \to \mathbb{R}$ such that

$$(Lf)|_S = 0,$$
$$f|_{\delta S} = g. \tag{A.2}$$

The generator can be considered as a linear operator from the space of real-valued functions on X to itself. These functions can then be mapped linearly to functions on S by the restriction operation, denoted R_S. In this language, the first equation above is equivalent to requiring that f lies in the kernel of $R_S L$. This kernel is equal to the kernel of $R_S \Delta$, where Δ is the matrix with elements

$$\Delta(x, y) = \begin{cases} 1 & x = y \\ -\dfrac{w_{xy}}{d_x} & x \neq y \end{cases} \quad \forall x, y \in X, \tag{A.3}$$

and we have defined $d_x = \sum_{y \in X} w_{xy}$. This matrix Δ is not symmetric, but its conjugacy class contains a symmetric matrix $\mathbb{L} = T^{1/2}\Delta T^{-1/2}$, where $T(x, x) = d_x$, and $T(x, y)$ is zero otherwise. The matrix $\mathbb{L}$ is called the normalized Laplacian of X. The most important object is the submatrix $\mathbb{L}_S$ of $\mathbb{L}$ consisting of matrix elements for which $x, y \in S$. This is a square symmetric matrix, and therefore, it has an orthonormal basis of eigenfunctions.

Let $\{(\phi_i, \lambda_i), i \in \mathcal{I}\}$ be such an orthonormal eigensystem of $\mathbb{L}_S$, and set $\tilde{f} = T^{1/2}f$. Observe now that for all $x \in S$:

$$(R_S \mathbb{L}\tilde{f})(x) = \sum_{y \in S \cup \delta S} \mathbb{L}(x, y)\tilde{f}(y) = \mathbb{L}_S \tilde{f}_S(x) + \sum_{y \in \delta S} \mathbb{L}(x, y)\tilde{f}(y) = 0. \tag{A.4}$$

So we need to solve the equation $\mathbb{L}_S R_S \tilde{f} = \alpha$, where

$$\alpha(x) = -\sum_{y \in \delta S} \mathbb{L}(x, y)\tilde{f}(y) = \sum_{y \in \delta S} d_x^{-1/2} w_{xy} g(y) \tag{A.5}$$

The matrix-tree theorem [31] states that the number of spanning trees of S equals $\det \mathbb{L}_S \times \prod_{x \in S} d_x$. Because S is connected, this number is nonzero, so $\mathbb{L}_S$ is invertible. Using that $\mathbb{L}_S = \sum_{i \in \mathcal{I}} \lambda_i \phi_i \phi_i^T$, we can compute

$$R_S f = T^{-1/2} R_S \tilde{f} = T^{-1/2} \mathbb{L}_S^{-1} \alpha = \sum_{i \in \mathcal{I}} \frac{1}{\lambda_i} T^{-1/2} \phi_i \phi_i^T \alpha. \tag{A.6}$$

If we work out this last expression, then we see that

$$f(x) = \sum_{i \in \mathcal{I}} \frac{1}{\lambda_i} \sum_{y \in S} \sum_{z \in \delta S} g(y) w_{yz} \phi_i(z) d_y^{-1/2} d_x^{-1/2} \phi_i(x) \tag{A.7}$$

for all $x \in S$.

Appendix B. Calculating the integral

The integral in equation (18) can be solved algebraically, and here we describe a method to do so. Starting from the integral

$$f(k, l) = \frac{1}{2\pi^2} \int_0^\pi \int_0^\pi \frac{\sin(u)\sin(v)\sin(ku)\sin(lv)}{(1 - \cos(u))(1 - \cos(u)/2 - \cos(v)/2)} \, du \, dv, \tag{B.1}$$

first simplify the integrand by introducing Chebyshev polynomials of the second kind, defined as $U_k(\cos(u)) = \frac{\sin((k+1)u)}{\sin(u)}$. This results in

$$f(k,l) = \frac{1}{2\pi^2} \int_0^\pi \int_0^\pi \frac{\sin(u)^2 \sin(v)^2 U_{k-1}(\cos(u)) U_{l-1}(\cos(v))}{(1-\cos(u))(1-\cos(u)/2-\cos(v)/2)} \mathrm{d}u\,\mathrm{d}v$$

$$= \frac{1}{2\pi^2} \int_0^\pi \int_0^\pi (1+\cos(u)) \frac{\sin(v)^2 U_{k-1}(\cos(u)) U_{l-1}(\cos(v))}{\sin^2(u/2) + \sin^2(v/2)} \mathrm{d}u\,\mathrm{d}v, \qquad (\text{B.2})$$

where the last line follows from some trigonometric identities. Then expand the polynomials, together with the factor $1 + \cos(u)$. This yields a number of terms, each of which is of the form

$$I_{mn} = \int_0^\pi \int_0^\pi \frac{\sin^2(v) \cos^m(u) \cos^n(v)}{\sin^2(u/2) + \sin^2(v/2)} \mathrm{d}u\,\mathrm{d}v \qquad (\text{B.3})$$

for some $m, n \in \mathbb{N}$. Those integrals can be transformed by changing variables to $x = u/2$ and $y = v/2$, and using the appropriate double angle formulas,

$$I_{mn} = 4 \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \frac{(\sin^2(x) - \sin^4(x))(1 - 2\sin^2(x))^m (1 - 2\sin^2(y))^n}{\sin^2(x) + \sin^2(y)} \mathrm{d}x\,\mathrm{d}y. \qquad (\text{B.4})$$

Again, expand the nominator and separate the terms, which now look like

$$J_{mn} = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \frac{\sin^{2m}(x) \sin^{2n}(y)}{\sin^2(x) + \sin^2(y)} \mathrm{d}x\,\mathrm{d}y. \qquad (\text{B.5})$$

Finally, use long division, noting that

$$\frac{x^m y^n}{x+y} = x^m y^{n-1} - x^{m+1} y^{n-2} + \ldots + (-1)^n \frac{x^{m+n}}{x+y}. \qquad (\text{B.6})$$

The resulting integrals can then be directly solved for all $m, n > 0$:

$$\int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \sin^{2m}(x) \sin^{2n}(x) \mathrm{d}x\,\mathrm{d}y = \frac{\pi^2}{2^{2m} 2^{2n}} \binom{2m-1}{m-1} \binom{2n-1}{n-1}, \qquad (\text{B.7})$$

$$\int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} \frac{\sin^{2m}(x)}{\sin^2(x) + \sin^2(y)} \mathrm{d}x\,\mathrm{d}y = \frac{\pi 2^n \Gamma(\frac{m}{2})^2}{16 \Gamma(m)}. \qquad (\text{B.8})$$

There is one curious fact we can derive from this calculation. The first integral above is always a rational number times π^2, and the second integral will be a rational times π if m is even and a rational times π^2 if m is odd. All the coefficients in the expansions are integers, and there is an overall factor of $1/\pi^2$, so the resulting integral can always be written as $a + b/\pi$, with $a, b \in \mathbb{Q}$.

This is not too unexpected, though. To derive the solution to a Dirichlet problem, one has to invert the Laplacian operator, in other words, derive Green's function. It is well known [32] that the two-dimensional lattice Green function is always of this particular form.

References

[1] Schiessel H 2003 The physics of chromatin *J. Phys.: Condens. Matter* **15** R699–774
[2] Luger K, Mäder A W, Richmond R K, Sargent D F and Richmond T J 1997 Crystal structure of the nucleosome core particle at 2.8 A resolution *Nature* **389** 251–60
[3] Maeshima K, Hihara S and Eltsov M 2010 Chromatin structure: does the 30-nm fibre exist *in vivo*? *Curr. Opin. Cell Biol.* **22** 291–7
[4] Polach K J and Widom J 1995 Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation *J. Mol. Biol.* **254** 130–49

[5]   Anderson J D and Widom J 2000 Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites *J. Mol. Biol.* **296** 979–87

[6]   Anderson J D and Widom J 2001 Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites *Mol. Cell. Biol.* **21** 3830–9

[7]   Anderson J D, Lowary P T and Widom J 2001 Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites *J. Mol. Biol.* **307** 977–85

[8]   Anderson J D, Thaström A and Widom J 2002 Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation *Mol. Cell. Biol.* **22** 7147–57

[9]   Li G, Levitus M, Bustamante C and Widom J 2005 Rapid spontaneous accessibility of nucleosomal DNA *Nature Struct. Mol. Biol.* **12** 46–53

[10] Kelbauskas L, Chan N, Bash R, Yodh J, Woodbury N and Lohr D 2007 Sequence-dependent nucleosome structure and stability variations detected by Förster resonance energy transfer *Biochemistry* **46** 2239–48

[11] Gansen A, Valeri A, Hauger F, Felekyan S, Kalinin S, Toth K, Langowski J and Seidel C A M 2009 Nucleosome disassembly intermediates characterized by single-molecule FRET *Proc. Natl Acad. Sci. USA* **106** 15308–13

[12] Koopmans W J A, Buning R, Schmidt T and van Noort J 2009 SpFRET using alternating excitation and FCS reveals progressive DNA unwrapping in nucleosomes *Biophys. J.* **97** 195–204

[13] Pennings S, Meersseman G and Bradbury E M 1991 Mobility of positioned nucleosomes on 5 S rDNA *J. Mol. Biol.* **220** 101–10

[14] Meersseman G, Pennings A and Bradbury E M 1992 Mobile nucleosomes—a general behavior *EMBO J.* **11** 2951–9

[15] Pennings S, Meersseman G and Bradbury E M 1994 Linker histones H1 and H5 prevent the mobility of positioned nucleosomes *Proc. Natl Acad. Sci. USA* **91** 10275–9

[16] Flaus A and Richmond T J 1998 Positioning and stability of nucleosomes on MMTV 3′LTR sequences *J. Mol. Biol.* **275** 427–41

[17] Gottesfeld J M, Belitsky J M, Melander C, Dervan P B and Luger K 2002 Blocking transcription through a nucleosome with synthetic DNA ligands *J. Mol. Biol.* **321** 249–63

[18] Pisano S, Marchioni E, Galati A, Mechelli R, Savino M and Cacchione S 2007 Telomeric nucleosomes are intrinsically mobile *J. Mol. Biol.* **369** 1153–62

[19] Prinsen P and Schiessel H 2010 Nucleosome stability and accessibility of its DNA to proteins *Biochimie* **92** 1722–8

[20] Kulic I M and Schiessel H 2003 Chromatin dynamics: nucleosomes go mobile through twist defects *Phys. Rev. Lett.* **91** 148103

[21] Chou T 2007 Peeling and sliding in nucleosome repositioning *Phys. Rev. Lett.* **99** 058105

[22] Mohammad-Rafiee, Kulic I M and Schiessel H 2004 Theory of nucleosome corkscrew sliding in the presence of synthetic DNA ligands *J. Mol. Biol.* **344** 47–58

[23] Blossey R and Schiessel H 2011 The dynamics of the nucleosome: thermal effects, external forces and ATP *FEBS J.* **278** 3619–32

[24] Chung F and Yau S T 2000 Discrete Green's functions *J. Comb. Theory* A **91** 191–214

[25] Fayolle G, Iasnogorodski R and Malyshev V 1999 *Random Walks in a Quarter Plane: Algebraic Methods, Boundary Value Problems and Applications* (Berlin: Springer)

[26] Menshikov M and Petritis D 2012 Explosion, implosion, and moments of passage times for continuous-time Markov chains : a semimartingale approach arXiv:1202.0952 [math.PR]

[27] Godréche C *et al* 1995 Spontaneous symmetry breaking: exact results for a biased random walk model of an exclusion process *J. Phys. A: Math. Gen.* **28** 6039

[28] Liggett T M 2005 *Interacting Particle Systems* (Berlin: Springer)

[29] Lamperti J 1977 *Stochastic Processes: A Survey of the Mathematical Theory* (Berlin: Springer)

[30] Kistler N *Stochastic Analysis II (Lecture Notes for the Graduate Course)* (Bonn: University of Bonn) (http://wiener.iam.uni-bonn.de/~nkistler/)

[31] Chung F R K 1997 *Spectral Graph Theory* (Providence, RI: American Mathematical Society)

[32] Spitzer F 1976 *Principles of Random Walk* 2nd edn (Berlin: Springer)