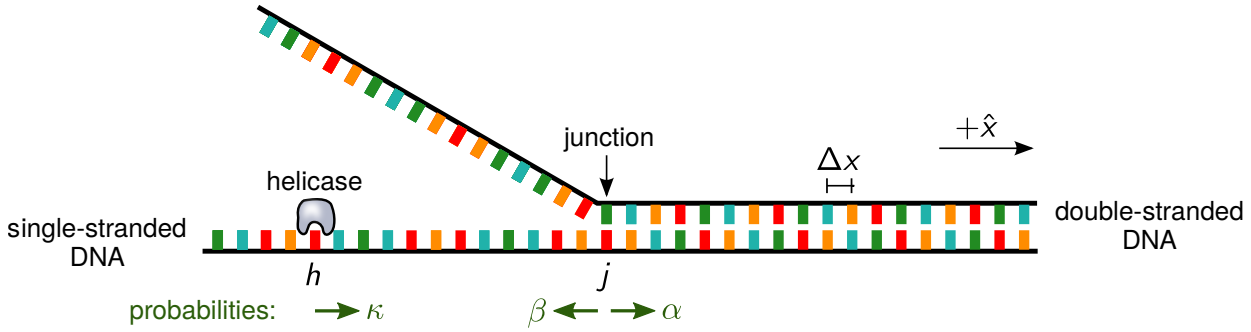
# PHYS 414 Problem Set 2: DNA, disease, driving

## Problem 1: Unwinding DNA




Helicases are proteins that can slide along single-stranded DNA, hopping between adjacent bases, mainly in one direction. They cannot move along double-stranded DNA, so a junction of two single strands, as shown above, is an obstacle to the helicase. The statistical physics of this protein can be described by a simple model, first introduced by Betterton and Jülicher [Phys. Rev. Lett. **91**, 258103 (2003)]. Imagine that one strand of the DNA is oriented along the $\hat{\mathbf{x}}$ axis, and denote the location of the helicase as $h\Delta x$, where $h$ is an integer, and $\Delta x$ is the distance between bases. The location of the junction (the first base pair in the double-stranded portion) is $j\Delta x$ for some integer $j$, where $j > h$. For a small time step $\delta t$, the probability of the helicase moving one base to the right ($h$ increasing by 1) is $\kappa$, but only if $j > h + 1$ (there is an open base of single-stranded DNA to the right of the helicase). If $j = h + 1$ the helicase cannot move. For simplicity we will assume the probability of helicase left steps is negligible. The probability of the junction moving to the right (a base pair unzipping, increasing $j$ by 1) is $\alpha$, while the opposite process (a base pair rezipping, decreasing $j$ by 1) has probability $\beta$. This rezipping cannot occur if the helicase is in the way, when $j = h+1$. Generally $\alpha < \beta$, so if the helicase is not obstructing ($j > h+1$) the junction will on average move left, rezipping the DNA with velocity $(\alpha-\beta)\Delta x/\delta t$. If the helicase is obstructing ($j = h+1$) then the junction will on average move right, unzipping the DNA with velocity $\alpha\Delta x/\delta t$. Similarly, if the helicase is not next to the junction ($j > h + 1$), the helicase on average moves to the right with velocity $\kappa\Delta x/\delta t$. If the helicase is next to the junction ($j = h + 1$) its velocity is zero.

**a)** Label the state of the system by the separation $n \equiv j - h = 1, 2, 3, \ldots$. Write down the transition matrix $W$ for this system. *Note:* The transition matrix is infinite, because $n$ can be arbitrarily large, so it is enough to write down the entries in the first few rows and columns of $W$, to see the pattern of the whole. You should find that the only nonzero entries in the matrix are either along the diagonal, or right next to the diagonal. In calculating the probability of seeing a transition $n \to n + 1$ or $n \to n - 1$ in time step $\delta t$, keep in mind the following assumptions: 1) All the different events described above ($h$ increasing by 1 or $j$ changing by $\pm 1$) are independent. Hence if a certain change in $n$ can happen in more than one way, the probabilities of those different ways add together. 2) To make life simple, we will assume that $\alpha$, $\beta$, and $\kappa$ are small enough that products of any two of them can be neglected. For example the probability that $h$ increases by 1 and $j$ decreases by 1 in the *same* time step is $\beta\kappa$, but we will assume this is $\ll 1$

and ignore it. Thus the entries in $W$ should be of linear order in $\alpha$, $\beta$, and $\kappa$, with all quadratic terms and higher neglected as an approximation. In any real physical system, we can generally choose $\delta t$ to be small enough for this approximation to be valid.

**b)** What does the graph corresponding to the matrix $W$ look like? Argue that it is microscopically reversible and hence ergodic. Thus as $t \to \infty$, we expect that $p_n(t) \to p_n^s$. To find the stationary distribution $\mathbf{p}^s$, look at corresponding eigenvalue equation:

$$W\mathbf{p}^s = \mathbf{p}^s$$

Though this is an infinite matrix multiplying an infinite vector, we can use the fact that most of the entries of $W$ are zero to make progress. Show that the eigenvalue equation leads to a set of recursion relations for the vector elements $p_n^s$, $n = 1, 2, \ldots$. Show that the solution of these recursion relations is $p_{n+1}^s = cp_n^s$, where $c$ is a constant. Find $c$ in terms of $\alpha$, $\beta$, and $\kappa$. Hence we know that $p_n^s = c^{n-1}p_1^s$. To complete the solution for the stationary distribution, we need to find the value of $p_1^s$, which we can do by demanding the stationary distribution be properly normalized: $\sum_{n=1}^{\infty} p_n^s = 1$. [Note that even when $p_n(t) = p_n^s$, the helicase and the junction are still moving: it is only the distribution of their separations that becomes time-independent.] Useful identity: $\sum_{k=1}^{\infty} x^k = x/(1-x)$ for $|x| < 1$.

**c)** Once the system has reached the stationary state, what is the overall mean velocity of the helicase? *Hint:* Based on the description in the introduction, consider how the velocity of the helicase depends on $n$. Then sum over the different possibilities, weighted by the probability $p_n^s$, to find the overall mean.

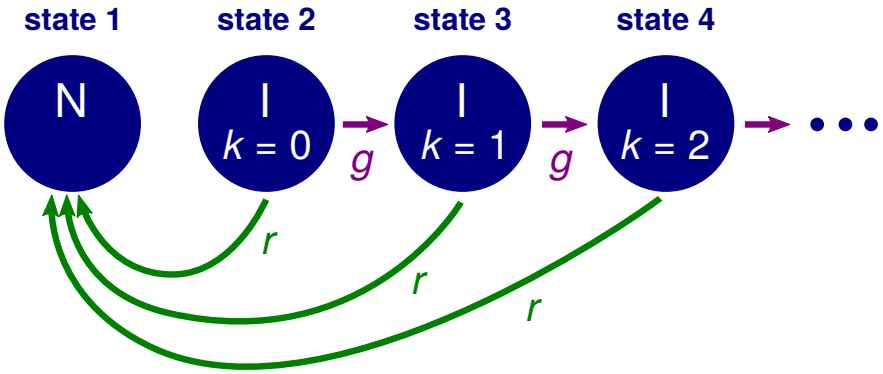
**d)** Once the system has reached the stationary state, what is the overall mean velocity of the junction? How does the mean velocity of the junction compare to that of the helicase from part c? Note that this relationship between the two velocities is a special feature of the fact that this is a *stationary* state.

**e)** The energy level of a DNA base pair in the unzipped state is $\epsilon > 0$ larger than the energy of the zipped state (because we have broken bonds to unzip). If there is one zipped and one unzipped conformation for the base pair, use the principle of local detailed balance to relate $\alpha$ and $\beta$ to $\epsilon$ and $T$ (the temperature of the environment). Note that $\beta$ in this problem is a probability, and not the inverse temperature. If we assume that $\beta \gg \kappa$, as is typically the case, what happens when $\epsilon$ gets larger (for example certain base pairs might have stronger interactions than others): does the helicase slow down or speed up? Does this result make physical sense?

## Problem 2: Disease spreading

We will consider a simple model of a single individual spreading a disease to a larger population. The individual may either be *infectious* (capable of spreading the disease) or *not infectious* (incapable of spreading the disease). Label the two conditions I and N. If we embed this individual in an arbitrarily large (infinite) healthy population, we can keep track of $k$, the number of people this individual has infected. Let state 1 correspond to the situation where the individual is N, regardless of what the value of $k$ may be (the individual's contagious period is over, whether because of recovery or death). Let state 2 correspond to the individual being I, with $k = 0$ (no spreading has occurred yet). State 3 is I and $k = 1$ (one other person has been infected by the individual). We continue the pattern of labeling states, and thus state $m$ for $m \geq 2$ corresponds to the individual being I and $k = m - 2$. Since the population is infinite, there are infinite possible states. Note we are only focusing on how many spreading events are due to our original contagious individual. We are not counting subsequent secondary infections in the population.

We consider two possible transition events that can occur at each time step $\delta t$. 1) With probability $r$ the individual can recover. The means that every state $m \geq 2$ can transition to state 1 with probability $r$. 2) If the individual is I, with probability $g$ one more person from the healthy population can get infected by the individual. This means that every state $m \geq 2$ can transition to state $m + 1$ with probability $g$. The kinetic network for this system is illustrated below:




Note that this network is *not* ergodic, but it does have a unique stationary state: if we wait long enough, we will eventually end up in state 1 with probability 1. We will be interested in particular about two aspects of the network dynamics.

The first question: if we start in state 2, how many people will get infected on average before the individual stops being infectious (transitions to state 1). In other words, how far on average will we travel to the right before returning to state 1. The answer to this question is an important number in epidemiology: the *basic reproduction number* $R_0$, the average number of new cases of infection that will be generated by the individual during a single round of the disease (counting only cases directly started by the individual). The common cold has an $R_0 \approx 6$, while for measles $R_0 \approx 15$, at the high end of the contagiousness scale among all known diseases. Ebola has $R_0 \approx 2$, and estimates for COVID-19 ranged from $2 - 4$ at the beginning of the pandemic. Food-borne illnesses like salmonella are on the low end of the spectrum, with $R_0 < 1$. The second question: what is $\bar{\tau}$, the average number of time steps the individual stay infectious?

In this problem, we will prove the following simple relations:

$$R_0 = \frac{g}{r}, \qquad \bar{\tau} = \frac{1}{r}.$$

As it turns, getting to these expressions will require some work. Intuitively, the result for $R_0$ just states that $R_0 = g\bar{\tau}$. In other words, the number of infections caused by the individual is the average number of time steps times the probability of spreading the infection per time step. Two sum identities will prove useful in this problem: i) $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ for $|x| < 1$; ii) $\sum_{n=0}^{\infty} nx^n = x/(1-x)^2$ for $|x| < 1$.

**a)** The first step is to write down the transition matrix $W$ for this system, corresponding to the graph above. As in Problem 1, the true matrix is infinite-dimensional, so you only need to write down a small portion of it to see the overall pattern. To keep the notation simple, you can introduce a new variable $\epsilon \equiv 1 - r - g$. This is useful for the diagonal entries, and corresponds to the probability that you stay in your current state at each time step.

**b)** Now let us consider all trajectories that involve infecting exactly $m$ people. These trajectories start at state 2 ($k = 0$) and reach state $m + 2$ ($k = m$), but no higher, before returning to state 1. We know that in general the probability of observing a trajectory of states $\nu = (n_0, n_1, \ldots, n_\tau)$ is given by $P(\nu) = W_{n_\tau n_{\tau-1}} \cdots W_{n_1 n_0} p_{n_0}(t_0)$. In our case all the trajectories start at $n_0 = 2$, so $p_{n_0}(t_0) = \delta_{n_0,2}$ is just a Kronecker delta function at $n_0 = 2$. We know that the final state $n_\tau = 1$. But there are a large number of possible trajectories, with different lengths $\tau$, that have these initial and final points and reach state $m + 2$. Imagine a trajectory $\nu$ in of this type with the following properties: you spend $a_2$ time steps in state 2 before transitioning to state 3; you spend $a_3$ time steps in state 3 before transitioning to state 4, and so on; finally, you spend $a_{m+2}$ time steps in state $m + 2$ before transitioning to state 1. The probability to stay in a state in each time step is $\epsilon$, the probability to transition to the next higher state is $g$, the probability to go from state $m + 2$ to 1 is $r$. Hence multiplying all these transition probabilities together, you get the following probability for the whole trajectory:

$$\mathcal{P}(\nu) = \epsilon^{a_2} g \epsilon^{a_3} g \cdots \epsilon^{a_{m+1}} g \epsilon^{a_{m+2}} r. \tag{1}$$

Let us call $\mathcal{P}^{(m)}$ the probability to infect exactly $m$ people. This is just the sum over all trajectories $\nu$ of the kind shown in Eq. (8), for all possible values of $a_2$, $a_3$, ..., $a_{m+2}$:

$$\mathcal{P}^{(m)} = \sum_{a_2=0}^{\infty} \sum_{a_3=0}^{\infty} \cdots \sum_{a_{m+2}=0}^{\infty} \mathcal{P}(\nu). \tag{2}$$

Evaluate this sum, and show that

$$\mathcal{P}^{(m)} = \frac{r g^m}{(r+g)^{m+1}}. \tag{3}$$

*Hint:* Do not forget the distributive property of sums, and take advantage of the sum identities mentioned above.

**c)** With the result of Eq. (10), calculating $R_0$ is now just a matter of evaluating the average of $m$ over all possible types of trajectories,

$$R_0 = \sum_{m=0}^{\infty} m\mathcal{P}^{(m)}. \tag{4}$$

Show that $R_0 = g/r$. *Hint:* manipulate the sum expression, factoring out certain terms, until you can apply one of the sum identities.

**d)** To calculate $\bar{\tau}$, note that $\bar{\tau} = h_{12}$, the *hitting time* $h_{12}$ starting from state 2 and ending in state 1. In class we derived a set of equations for the hitting times $h_{ji}$:

$$h_{ji} = 1 + \sum_{n \neq j} h_{jn} W_{ni}. \tag{5}$$

By plugging in the transition probabilities from part a, use Eq. (5) to write down a set of equations for $h_{12}$, $h_{13}$, $h_{14}$, and so on. You should find that $h_{12}$ depends on $h_{13}$, $h_{13}$ depends on $h_{14}$, etc. If you solve for $h_{12}$, and then plug in the $h_{13}$ solution, and then plug in the $h_{14}$ solution, etc., show that you get an infinite series of the form:

$$h_{12} = \frac{1}{r+g}\left[1 + \frac{g}{r+g} + \left(\frac{g}{r+g}\right)^2 + \left(\frac{g}{r+g}\right)^3 + \cdots\right]. \tag{6}$$

Evaluate the sum in the brackets, and show that the final expression you get is $h_{12} = 1/r$.

## Problem 3: Master equation under periodic driving

So far in class we have focused on time-independent transition matrices $W$, which correspond to systems where the environmental parameters are fixed. In this case the master equation is given by

$$\mathbf{p}(t+\delta t) = W\mathbf{p}(t). \tag{7}$$

Iterating this equation, for example over $m$ time steps, we get

$$\mathbf{p}(t+m\delta t) = W^m\mathbf{p}(t). \tag{8}$$

Assuming an ergodic and mixing system, as $m \to \infty$ the probability distribution converges to a stationary state, $\lim_{m\to\infty} \mathbf{p}(t+m\delta t) = \mathbf{p}^s$, where $\mathbf{p}^s$ is a time-independent vector.

But what happens if the environmental parameters are not fixed, so that the transition matrix probabilities change at every time step? Incorporating this into the master equation framework is straightforward: the transition probabilities $W_{ij}$ between states $j$ and $i$ in the system become time-dependent functions $W_{ij}(t)$, which could for example reflect some protocol of parameter changes controlled by the experimentalist. Note that at each instant in time the probabilities of each transition out of $j$ over the interval $\delta t$ have to still sum to one, $\sum_i W_{ij}(t) = 1$. The corresponding master equation becomes

$$\mathbf{p}(t+\delta t) = W(t)\mathbf{p}(t). \tag{9}$$

For arbitrary transition probabilities $W_{ij}(t)$, this master equation is not easy to manipulate, but in this problem we will be interested in a specific class of external stimulus which has the form of a cycle: in other words, $W_{ij}(t)$ is periodic, $W_{ij}(t+T) = W_{ij}(t)$, with some period $T = \tau\delta t$, where $\tau > 0$ is an integer. We will show that an ergodic system under these circumstances will go to a *periodic state* in the long-time limit: $\lim_{m\to\infty} \mathbf{p}(t+m\delta t) = \mathbf{p}^{ps}(t+m\delta t)$, where $p_n^{ps}(t+T) = p_n^{ps}(t)$ for any $n$. The nonequilibrium behavior of periodically driven systems is currently a very active research area in condensed matter, falling under the broader mathematical framework of Floquet theory. Interest in this topic was stimulated in particular by the suggestion by Frank Wilczek in 2012 that in certain quantum and classical systems, you can drive with period $T$ and get the system to be periodic with a period different than the driving, for example $2T$. Wilczek colorfully named this phenomenon a *time crystal*. Unfortunately in our case we will show that a Markovian, ergodic system under periodic driving is not an exotic time crystal: the periodic state has the same period $T$ as the the driving. But the fact that we are guaranteed to eventually reach this periodic state is still an interesting result, which we will exploit in the next problem set.

**a)** Let us first focus on times $t$ that are integer multiples of the period, $t = jT$ for $j = 0, 1, 2, \ldots$. By iterating Eq. (9) and using the periodicity $W(t+T) = W(t)$, show that $\mathbf{p}(jT+T)$ can be related to $\mathbf{p}(jT)$ as:

$$\mathbf{p}(jT+T) = \hat{W}\mathbf{p}(jT), \tag{10}$$

where $\hat{W}$ is a matrix independent of $j$. Find $\hat{W}$ in terms of products of the $W(t)$ matrices, and show that $\hat{W}$ has the standard property of a transition matrix, namely $\sum_m \hat{W}_{mn} = 1$.

Note that Eq. (10) has the same form as Eq. (7), with $t = jT$, $\delta t$ replaced by $T$ and $W$ replaced by $\hat{W}$. Thus it describes time evolution of the probability $\mathbf{p}$ over time intervals equal to $T$. We

will assume $\hat{W}$ corresponds to an ergodic network, so that as $t \to \infty$ (or equivalently $j \to \infty$) the probability $\mathbf{p}(jT) \to \mathbf{v}$, where $\mathbf{v}$ is some time-independent vector.

**b)** The final step is to find $\mathbf{p}(t)$ for some time $t$ that is an integer multiple of $\delta t$, but not necessarily an integer multiple of $T$. In other words, $t = k\delta t + \lfloor t/T \rfloor T$, for some $k$. Here $\lfloor x \rfloor$ is the floor of $x$ (the largest integer smaller or equal to $x$), and hence $\lfloor t/T \rfloor T$ is just the largest multiple of $T$ less than $t$. The remainder is $k\delta t = t - \lfloor t/T \rfloor T \equiv t \bmod T$. Using Eqs. (9) and (10), show that you can write:

$$\mathbf{p}(t) = A(t \bmod T)\mathbf{p}\left(\lfloor t/T \rfloor T\right), \tag{11}$$

where $A(t \bmod T)$ is a matrix depending on $t \bmod T$. Find an expression for $A(t \bmod T)$, and argue that as $t \to \infty$, the probability becomes periodic in $T$, namely $\mathbf{p}(t+T) = \mathbf{p}(t)$. Thus we have shown that the system must go to a periodic state at long times.