BE/Bi 101: Order-of-Magnitude Biology
Homework 2
Due date: Friday, January 23, 2015

> "Probability theory is nothing but common sense reduced to calculation."
>
> —Pierre-Simon Laplace

1. Using probability to figure out how molecular motors walk.

Kinesin is a motor protein that "walks" in a directed fashion along microtubules. A little more than a decade ago, it was unknown whether kinesin walked along microtubules hand-over-hand or like an inchworm (See Fig. 1).

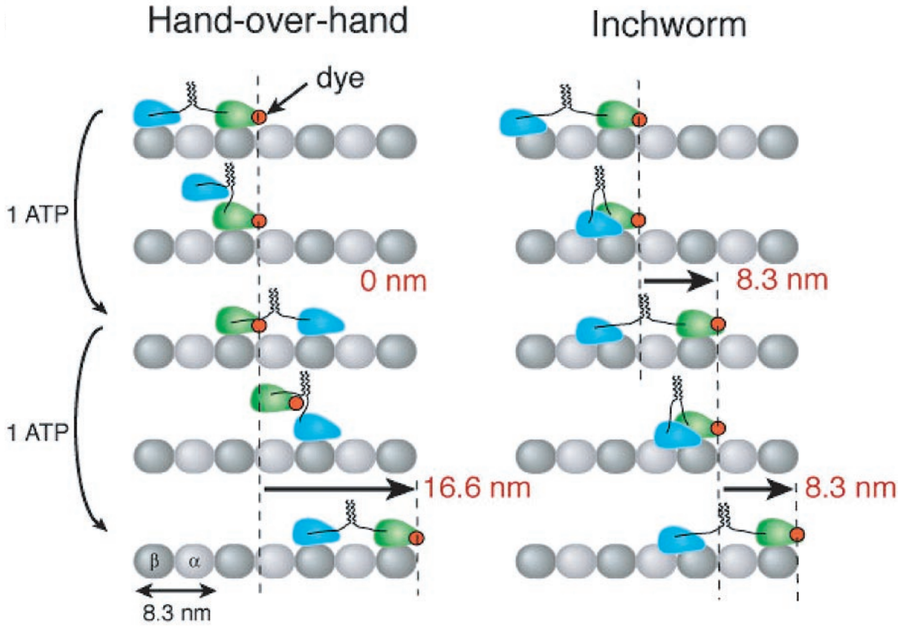

Figure 1: Schematic of hand-over-hand and inchworm mechanisms of kinesin walking on microtubules. Kinesin heads bind to the β subunit of the tubulin dimers that comprise a microtubule. Thus, the microtubule geometry dictates kinesin step size that is an integer multiple of the spacing between β subunits, 8.3 nm. Adapted from Yildiz, et al., *Science*, **303**, 676–678, 2004.

To settle this question, Yildiz and coworkers performed an elegant experiment. They tagged one of the "heads" of kinesin with a fluorescent dye. They then used total internal reflection fluorescence microscopy (TIRF) to monitor the movement of the fluorescent dye over time.[1] By performing a curve fit of the fluorescent signal in their images with a Gaussian (which approximates the point spread function of the fluorophore), they can pinpoint the position of

[1] You can see a movie of one of their fluorescent dyes, corresponding to the top trace of Fig. 3, here. The interpixel spacing is 86.6 nm. The movie lasts 20 seconds in total, with a frame rate of 3 frames per second.

the fluorescent tag to an accuracy close to one nanometer. Such a curve fit is shown in Fig. 2.

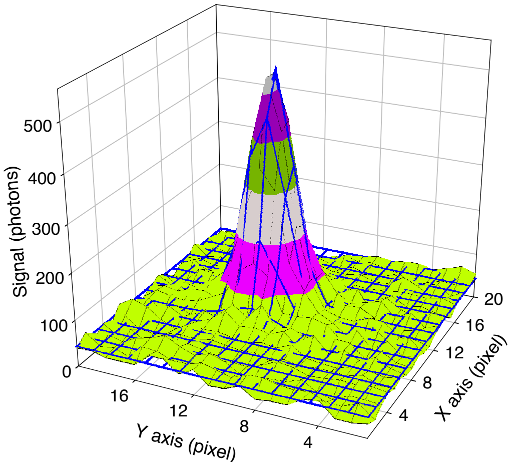

Figure 2: Plot of photon counts (colored surface) with Gaussian fit (blue mesh). The position of the maximum of the fitted Gaussian is used to pinpoint the position of the fluorophore. Figure taken from Yildiz, et al., *Science*, **303**, 676–678, 2004.

The dynamics of the motor is revealed in traces like those shown in Fig. 3. From these traces, they can compute the step length (the height of the vertical segments of the red lines) and the dwell time of the labeled head between movements (the length of the horizontal segments of the red lines).

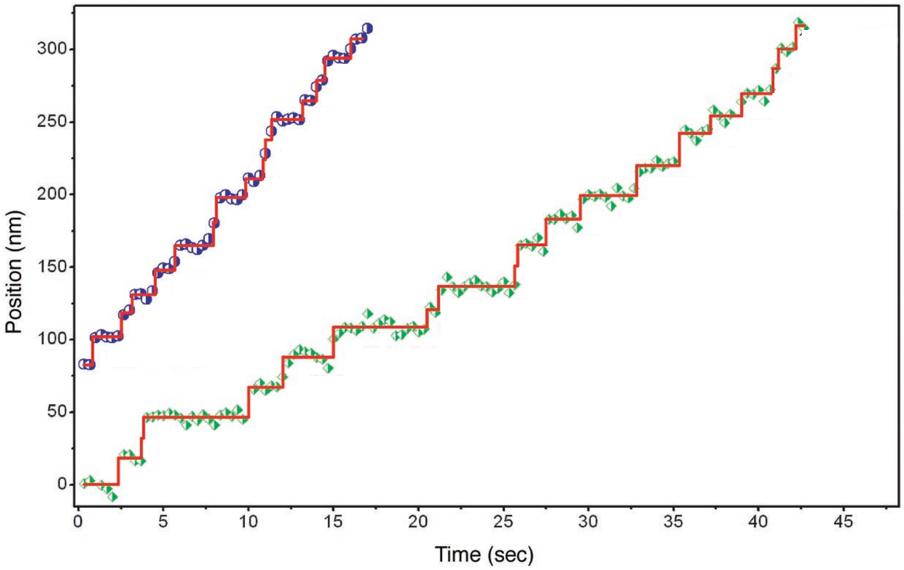

Figure 3: Sample traces of the position of the fluorophore versus time. Symbols are measurements; red lines are the results of data analysis. The steps (vertical red lines) are typically faster than the 0.5 s frame rate of the experiment, so they appear instantaneous. These traces are for homodimer kinesins (with only one of the feet fluorescently labeled). Adapted from Yildiz, et al., *Science*, **303**, 676–678, 2004.

a) Explain why Fig. 4A suggests that the kinesin more likely walks hand-over-hand than like an inchworm.

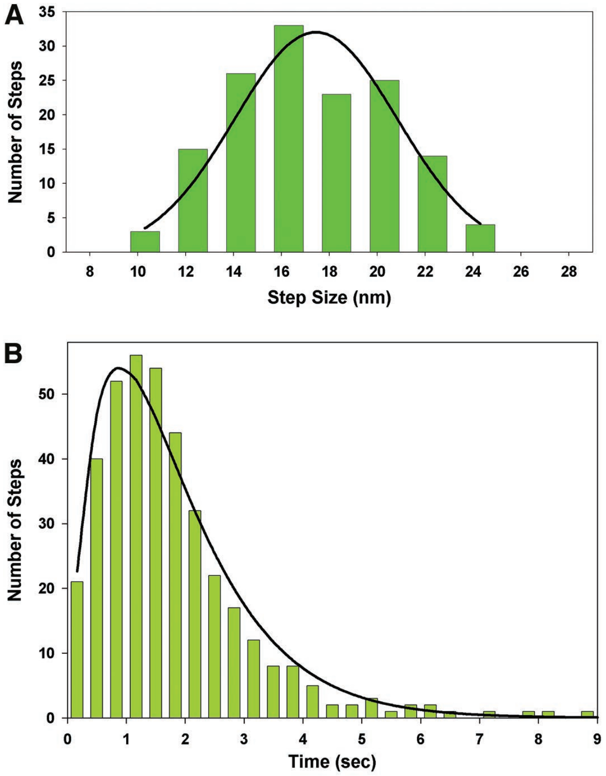

Figure 4: Experimentally determined histograms of A) fluorophore displacement distances and B) dwell times between displacements. From Yildiz, et al., *Science*, **303**, 676–678, 2004.

b) In the schematic of hand-over-hand motion shown in Fig. 1, each step consists of detachment of the back foot and its subsequent reattachment in the front position. Which do you think happens more rapidly, detachment or attachment of the head? Explain your reasoning.

c) To provide further evidence in favor of the hand-over-hand model to describe kinesin walking, we can investigate the dwell time of the fluorophore between displacements. We will first calculate how long we expect to wait for a single step. (In the next part, we will calculate how long we expect to wait to observe displacement of the fluorophore, which is equivalent to how long we have to wait for two kinesin steps.)

To do the calculation, we remind ourselves that each kinesin step consists of two processes, the detachment of the rear head and its re-attachment to the microtubule. Each of these two processes has a probability distribution that describes how long it will take for the respective event to happen. Our goal is to work out the probability distribution for the waiting time for detachment and re-attachment to *both* happen in succession.

i) Imagine a kinesin motor is in the configuration shown in the top left of Fig. 1. The first event to happen in taking a step is the detachment of the blue kinesin head from the microtubule. Assume that the probability distribution, $P_d(t)$, for the amount of time we have to wait for the detachment event is

$$P_d(t) = \tau_d^{-1} \mathrm{e}^{-t/\tau_d}. \tag{1}$$

Why is this an appropriate probability distribution?

ii) The next step involves the re-attachment of the blue head. We assume that the probability distribution for the waiting time of this event, $P_a(t)$, is also exponential,

$$P_a(t) = \tau_a^{-1} \mathrm{e}^{-t/\tau_a}. \tag{2}$$

Now, we want to know the probability distribution, $P_{\text{step}}(t)$, for the time that we have to wait for these two events to happen in succession. For this calculation, we assume that we start our stopwatch at time 0, and the reattachment, i.e. the completion of the step, happens at time t. The detachment would have to happen at time t_d with $0 \le t_d \le t$. To compute $P_{\text{step}}(t)$, we need to *marginalize* over the intermediate time t_d; i.e., we need to integrate over all possible times t_d. So, we have

$$P_{\text{step}}(t) = \int_0^t \mathrm{d}t_d\, P_d(t_d) P_a(t - t_d). \tag{3}$$

Explain why this is this right expression for $P_{\text{step}}(t)$.

iii) Perform the integral in equation (3) to show that

$$P_{\text{step}}(t) = \frac{\mathrm{e}^{-t/\tau_d} - \mathrm{e}^{-t/\tau_a}}{\tau_d - \tau_a}. \tag{4}$$

iv) Show that if $\tau_d \gg \tau_a$, $P_{\text{step}}(t)$ is approximately exponential for times significantly greater than τ_a. More generally, this means that if two independent events have to happen and if one is much slower, the dynamics are dominated by the slow one on long time scales. This is a useful bit of information to keep in your back pocket while making order-of-magnitude estimates of the rates of things.

d) We have shown in part (b) that we expect the waiting time for each kinesin step is exponentially distributed. Show that we would expect the waiting time for fluorophore displacements to be distributed as

$$P(t) = \frac{t}{\tau_d^2}\, \mathrm{e}^{-t/\tau_d}. \tag{5}$$

Sketch this function for various values of τ_d. *Hint*: You already derived an expression for two successive processes, each with exponential waiting times. Take the limit where the mean waiting times of the two processes are equal. You might need to use L'Hôpital's rule.

e) The experimental result for $P(t)$ is shown in Fig. 4b. Does this result lend credence to the hand-over-hand mode of walking?

After completing this problem, it is well worth reading the paper, which you can download here.

2. Probability and shotgun sequencing.

Shotgun sequencing is a widely used technique for genome sequencing. Such a technique is necessary because there is currently no technique to simply read a very long piece of DNA from its start until its end. Typically, we can only get sequences of about 500 nucleotides. A small sequenced region like this is called a *read*.

The idea behind shotgun sequencing is to take multiple copies of the DNA to be sequenced and then randomly break it up into small pieces. These small pieces are then sequenced, giving reads of length L nucleotides. The length of these reads is dependent on the sequencing method, ranging from 50 nucleotides to 10,000 for single-molecule real-time methods. For simplicity in this problem, we will assume L is the same for all reads.

The reads have some overlap, and sequence alignment algorithms are used to detect the overlaps and stitch together the genome. There can still be some gaps where there was no sequence overlap, so the entire genome will not be completely sequenced. A continuous region of the genome that has overlapping reads is called a *contig*. A nucleotide that is sequenced is then said to be in a contig and one that is not is in a *gap*.

We would like to have some rules of thumb about how many reads we have to do in order to get good information about an entire genome. We'll define the following variables (in their traditional nomenclature) in our analysis, namely,

$$\begin{aligned}
G &= \text{genome length,} \\
N &= \text{number of reads,} \\
L &= \text{read length,} \\
T &= \text{nucleotides of overlap needed to detect overlap.}
\end{aligned}$$

a) Consider some specific interval in the genome consisting of L consecutive nucleotides. What is the probability $P(n; N, G, L)$ that n reads start within the interval? *Note on notation*: The semicolon in the function arguments is meant to separate the variable over which the distribution is considered (n) and the parameters (N, G, L). So, your expression for the probability distribution should have dependence on these parameters. *Hint*: Can you recast this problem into a "story" that corresponds to one of the probability distributions derived in class[2]? Is there then a convenient limit you can take?

[2] By "story" we mean the stories that describe a certain probability distribution. For example, the exponential distribution describes the waiting time for an event to happen. The binomial distribution describes how many heads you get in a series of coin flips of a (possibly biased) coin. Etcetera.

b) How much of the genome is contained in gaps? In other words, how many nucleotides are not sequenced?

c) If we are sequencing the human genome with shotgun sequencing with a read length of $L \approx 500$ nucleotides (as is typical of Sanger or 454 methods) how many reads N would be need to do to have 99.9% of the genome sequenced? How many total sequenced bases is this? Is there enough RAM in your computer to store all of these? How many bases in the genome remain unsequenced?

d) (5 pts extra credit) What is the probability distribution $P(n_{\text{reads}}; N, G, L)$ describing the number of reads in a given contig, n_{reads}? For this rough estimate, assume $T/L \approx 0$. There is also some variation in the literature concerning the definition of a contig; some require at least two reads. We will define a contig to be any subsequence of the genome that has at least one read.

e) (5 pts extra credit) What is the probability distribution, $P(n_{\text{contig}}; N, G, L)$, for the number of contigs? What is the expected number of contigs?