Programmatically locating the cutting point of circular DNA in nanochannels

Charleston Noble

Figure 1: Steps in detecting the circular DNA region. From left to right: (a) The raw kymograph. (b) Smoothing. (c) Isolating the circular DNA region. (d) Trimming the circular region. (e) Identifying the cutting point (red).

The method

Given the kymograph of a circular DNA molecule stretched in a nanochannel (Fig. 1a), we aim to find its cutting time and the position on the molecule where this cutting occurs. On our kymograph, these values correspond to the vertical and horizontal coordinates, respectively, where the upper light region first transitions to the lower darker region (Fig. 1e).

Our method relies on the assumption that the circular DNA, linear DNA, and background regions all have consistent and easily-differentiated intensity values. From this assumption, we can isolate the circular DNA region (Fig. 1c) and locate the cutting point. This is done through the following process (details can be found in Appendix A).

1. We begin by smoothing the raw kymograph using a 2D moving average (Fig. 1b). This eases identification of the three types (circular, linear and background) of intensity values in Step 2. See Fig. 2 for an illustration.

2. The image is segmented into three regions corresponding to circular DNA, linear DNA, and background. The circular region is isolated (Fig. 1c).

3. The cutting time is estimated, and the circular region's edges above this point are trimmed to better define the length of the molecule (Fig. 1d).

4. The horizontal cutting location is identified as the column with the smallest nonzero white pixel count. The cutting time is the number of rows before the last white pixel is found in this column.

Appendix A: Details of the method

Step 1

The segmentation algorithm discussed below in Step 2 uses the kymograph's intensity histogram (Fig. 2a) to differentiate between our regions of interest. In the best case scenario, this histogram would contain three clear peaks corresponding to our regions (circular DNA, linear DNA, and background). However, random intensity fluctuations introduce normally-distributed background noise which obscures the peaks. To obtain more definition, a moving average is performed in both dimensions to decrease peak overlap (Figs. 1b and 2b). In the vertical direction, we use a 7-pixel window, and in the horizontal direction we use a 3-pixel window. The moving average was chosen over other smoothing methods because it is optimal in uncovering underlying trends when fluctuations about the trend are normally distributed.[1]

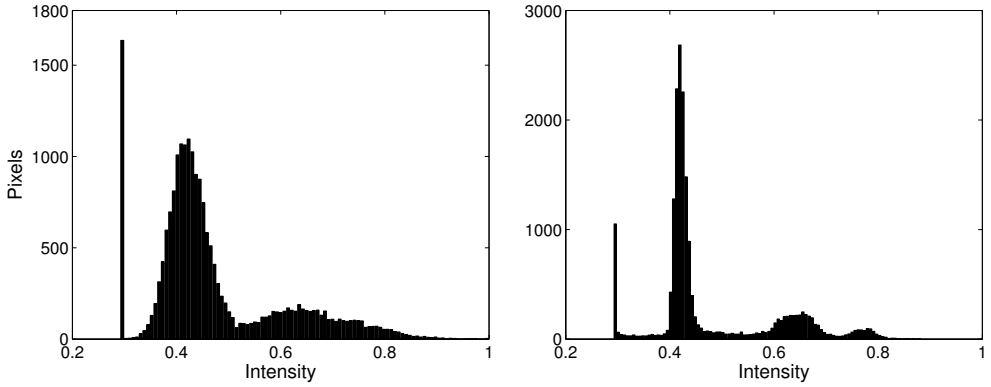

Figure 2: Intensity histograms of (a) the raw kymograph, and (b) the kymograph after smoothing. Note that the smoothing process narrows the intensity peaks and creates definition between the circular and linear DNA regions (far right peak and second from right, respectively).

Step 2

For this segmentation process, we use the multi-Otsu approach.[2] Intuitively, this method discretizes pixel intensity values by assigning thresholds. Pixels are assigned to classes based on these thresholds (i.e., two pixels are assigned to the same class if their intensities fall between the same threshold values). The Otsu approach assigns threshold values which globally minimize intra-class intensity variance in the following way: consider our kymograph as a 2D grayscale intensity function, containing N pixels. Then we can define the probability p_i of a pixel having gray level i as

$$p_i = f_i/N$$

where f_i is the number of pixels with gray level i. The collection of these p_i's is essentially the intensity histogram of our smoothed kymograph, normalized by its pixel count (Fig. 2b). Next, we compute the *within-class* intensity variance for thresholds placed at gray levels T_1 and T_2 (with $T_1 < T_2$). This is defined as

$$\sigma^2_{\text{within}}(T_1, T_2) = n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2$$

where σ_1^2, σ_2^2, and σ_3^2 denote the intensity variances for pixels falling in each of the three classes, and n_i represents the sum of the ith bucket's p_i values. This is essentially the sum of class variances weighted by class size. T_1 and T_2 are then optimized by exhaustive search to minimize this within-class variance, and our three resulting classes correspond to the background region, the linear DNA region, and the circular DNA region, respectively. In some cases, the signal-to-noise ratio is high enough that the entire DNA-containing

[1] G.R. Arce, "Nonlinear Signal Processing: A Statistical Approach", Wiley:New Jersey, USA, 2005.
[2] Journal of Information Science and Engineering 17, 713-727 (2001)

region is placed into the same class, and in this event we simply perform the segmentation process again, except we only consider the DNA-containing region and seek only one threshold. After this process is complete, we create a mask using the largest connected component of pixels falling in the circular DNA class (Fig. 1c).

Step 3

After the circular region has been identified, we estimate the time at which cutting occurs. For this step, we sum the masked image horizontally, giving us the linear length of the circular DNA at each time step. We denote this by $S_H(t)$ (Fig. 3). We then define a measure $\sigma_S^2(t)$ which is the variance of S_H for times after and including t, or

$$\sigma_S^2(t) = \mathrm{Var}\left[S_H(k \geq t)\right]$$

Because of the characteristic shape of $S_H(t)$, which stays somewhat constant until a rapid drop-off at the cutting point, our $\sigma_S^2(t)$ measure is maximized just before the DNA cuts, providing us with a reliable estimate of the cutting time (Fig. 3).

Next we smooth the edges of the masked region to better define its boundaries. This is important for determining the relative position of the cut on the DNA molecule itself. To accomplish this, we simply eliminate columns which contain one or more non-masked pixels for t less than the cutting time found above (Fig. 1d).

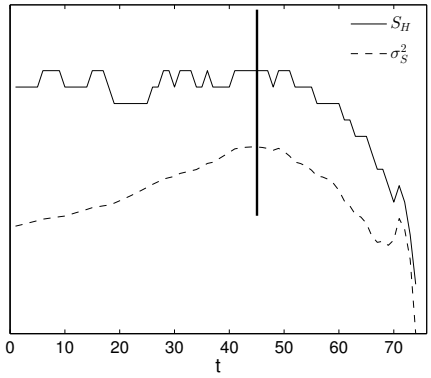

Figure 3: Our method for estimating the DNA cutting time. S_H is just the horizontal sum of the binary mask for the circular DNA region. This measure always remains somewhat constant until the DNA cuts, and then the circular region narrows rapidly. Thus our $\sigma_S^2(t) = \mathrm{Var}(S_H(t^* \geq t))$ measure is maximized just before cutting. The vertical bar represents the estimated cutting time.

Step 4

We then sum the masked image vertically, giving us the total number of time frames including circular DNA for a given coordinate, x. We denote this $S_V(x)$. We take the mean of x-coordinate(s) that minimize S_V and label this the absolute cutting coordinate. The absolute cutting time is then the largest time for which this column contains a pixel in the circular DNA class. Finally, the relative cutting position is simply the number of columns containing circular DNA to the left of this point divided by the absolute width of circular DNA determined in Step 3.

Appendix B: Results and Discussion

We validated our method on approximately 250 kymographs, and the results are shown below. The original kymograph is shown with the cutting point marked, as well as the isolated circular DNA region for reference. Cutting points assigned by our method seemed consistent with visually-placed cutting points for over 90% of our data. The algorithm seems to perform best when there is high contrast between the circular DNA, linear DNA, and background regions, and when the time axis is centered around the breaking point.

The method performs poorly when the DNA-containing region is dominated by either circular or linear DNA–that is, when the cutting point occurs very near the beginning or end of the time-scale (for example, see the bottom right figure on page 19). In these cases, the intensity histogram loses the peak corresponding to the region that was under-represented, and the thresholds cannot be placed properly by Otsu's algorithm. Thus the circular and linear DNA regions cannot be properly segmented. Similarly, our method could be improved for situations with low contrast between circular and linear DNA regions, regardless of the cut location on the time-axis (for example, see the top left figure on page 30). The algorithm fails in this situation because the circular DNA region cannot be properly segmented due to the low contrast between the regions.

Fortunately, as we've now seen, the algorithm generally fails only during the segmentation phase. By the nature of Otsu's method, which minimizes intra-class intensity variance, we could easily define a natural "confidence" measure based on the minimum variance and return this with the results.

All that said, the approach does work very well when given experimental data meeting the criteria above. And furthermore, it seems fairly robust to the level of background noise (for example, see page 20 for figures with varying noise levels). Especially with the incorporation of our confidence measure, the method will work quite well for cut identification.