pBR322 Mapping by Restriction Digestion with EcoR1/HincII/PvuII and Electrophoresis

Xun Zhao (Partner: Samantha)

March 15, 2019

Introduction

In order to defend the attack from virus, some bacterias can produce enzymes to digest invasive DNA molecule of virus. Among these kinds of enzymes, some of them cut DNA randomly, while others called restriction enzymes only cut at specific sequence. Thus, we can use these enzymes to digest the plasmid and get DNA fragments with different length.

As the cut sites are fixed in a plasmid, we can reconstruct the relative positions of restriction sites of a plasmid, called mapping.

Firstly, we digest the plasmid with all possible combinations of 3 restriction enzymes. Then, we can use the agarose gel and certain voltage to separate negatively charged DNA fragments and use loading dye to indicate their positions under UV. And their travel distances are proportional to the reciprocal of logarithm of the number of base pairs, namely, the size of molecule compared with the size of holes in gel, which allows us to draw a standard curve from a reference plasmid with known structure. Finally, we can get the length from travel distance based on the standard curve and use these length to rebuild a plasmid.

1 Results

1.1 Raw Result Photo

To make the photo more readable, I convert[1] the photo to black-and-white, adjust the contrast and label 9 wells.

Figure 1: Raw Gel Photo

1.2 λ DNA Data

Here is the known information of λ DNA's fragments bands digested by **HindIII**, which is the 8^{th} well in the photo named **L**.

Table 1: λ DNA Data

	base pairs	log(bps)	distance (pixels)
1	23130	4.364175633	266
2	9461	3.97386645	292
3	6557	3.816705184	326
4	4361	3.639586087	376
5	2322	3.365862215	468
6	2027	3.306853749	502
7	564	2.751279104	870

1.3 Standard Curve

We can write distance in terms of d, and base pairs in terms of n. According to the relationship, we will have,

$$d \propto \frac{1}{log(n)}$$

The raw curve is plotted as follows,

Figure 2: The realtionship between distance and logarithm of number of base pairs

After choosing some of the data points and using linear regression fitting algorithm, I derived the equation that:

$$n = exp(4.2 - 1.7 \times 10^{-3} \cdot d)$$

where n is the length (bps) of DNA fragments and d is the distance these fragments travel.

[1] All the picture analysis below is based on the *Fiji ImageJ*, including color converting and distance measuring.

The combination of this line and data points is as follows,

Figure 3: Standard Curve with Fitting Line

1.4 pBR322 Digestion Data

The distance of every well is shown below, using pixel unit.

Table 2: Distance

Tube	E	H	P	EH	EP	HP	EHP
Distance (pixels)	400.125	418.12 682.237	388.082	424.118 676.027 830.039 924.078	472.004 486.037	520.004 596.03 668.012	530.015 602.003 830.002 928.002

So using the equation derived above, we can get the estimated sequence length (rounded to integer),

Table 3: Length

Tube	E	H	P	EH	EP	HP	EHP
Length (bps)	3342	3112 1107	3499	3041 1135 621 430	2523 2388	2089 1552 1169	2009 1517 621 423

and also the logarithm of length,

Table 4: Logarithm of Length

Tube	E	H	P	EH	EP	HP	EHP
log(Length)	3.524	3.493 3.044	3.544	3.483 3.055 2.793 2.633	3.402 3.378	3.32 3.191 3.068	3.303 3.181 2.793 2.626

1.5 pBR322 Mapping

Figure 4: Mapping Result

2 Discussion

2.1 Overview of Discussion

First, we will fit the standard curve.

Next, as the random error is inevitable in experiment, there are hardly two bands in photo that have the same distance in pixels. We have to assume that some of them indeed traveled the same distance.

Then, we will analyze the relative position without exact length information. Instead, the analysis will be based on the assumption that some bands seem to be same.

Finally, we will using statistical method to estimate the exact length, using the original data to minimize the estimation error.

2.2 Standard Curve

First, I tried to discard the points with larger length, which is out of the gel's resolving ability. The lines are derived from linear regression, namely, least squared method (calculation not shown).

Figure 5: No Discarding

Figure 6: Discarding Point 1

Figure 7: Discarding Point 1∼2

Figure 8: Discarding Point 1∼3

Then compared the R^2, we can see that discarding first 3 points leads to the highest R^2 as 0.976. In addition, these three points are also longer than 5000bp, larger than the ability of gel resolution.

If we discard more points, we would get higher R^2, but we will lose data information and increase the effect of random error.

Finally we will use Figure 8 as the fitting line, which has already been plotted in Figure 3 in **Result** part.

In conclusion,

$$\log(n) = 4.2 - 1.7 \times 10^{-3} \cdot d$$
$$n = exp(4.2 - 1.7 \times 10^{-3} \cdot d)$$

2.3 Relative Position Analysis

For convenience, I label bands from tube **E, H, P, EH, EP, HP, EHP**.

Figure 9: Labeled Bands

2.3.1 Single Enzyme Digestion

For tube **E, H, P**, we can see that **E, P** have two similar single bands, but **H** has two bands, which indicates that enzyme EcoR1 and PvuII cut the plasmid at one site, while HincII cuts at two sites. The picture below demonstrates this more clearly.

Figure 10: Number of Cutting Sites

Here, I choose different lengths for H.1 and H.2, because their bands' distances are different, causing H.1 to be longer with small distance, and H.2 to be shorter.

$$\text{pBR322} \underset{\text{HincII}}{\rightarrow} \text{H.1} + \text{H.2}$$

$$\text{H.1} > \text{H.2}$$

2.3.2 Two Enzymes Digestion

After we get the number of cutting sites of three enzymes, we can analysis their combination digestion.

EH tube There are 4 different bands, and two of them (EH.1, EH.2) are almost the same as bands of **H**'s (H.1, H.2). However, as it is digested by EcoR1 and HincII, who have 1 and 2 cutting sites. So there should be 3 sites, and 3 bands. That is, HincII cuts the plasmid into two part, and EcoR1 cuts one of these two in to another two parts. Considering two same bands as **H**'s, the reaction of EcoR1's cutting is not complete, remaining small amount of HincII's product. In addition, as the signal of EH.1 is stronger than EH.2's, it is very likely that EH.2 (also H.2) should be cut into EH.3 and EH.4 but reaction is not complete, which means EcoR1's site might be located on EH.2 (also H.2), namely, shorter part between two HincII sites.

$$\text{H.2} \underset{\text{EcoR1}}{\rightarrow} \text{EH.3} + \text{EH.4 (highly likely)}$$

EP tube The phenomenon is similar to **H**'s. Simply put, EcoR1's and PvuII's cutting sites lead to two fragments with different but similar lengths, since their distances are similar. So, EcoR1's cutting site should be about the opposite of PvuII's site.

$$\text{pBR322} \underset{\text{EcoR1, PvuII}}{\rightarrow} \text{EP.1} + \text{EP.2 (similar length)}$$

HP tube This can be analyzed as **EH** tube, but with better and clearer result. The digestion process can be explained as follows[2]. HincII cuts plasmid into H.1 and H.2. Then, PvuII cuts H.1 into HP.1 and HP.2, leaving H.2 as HP.3, for their lengths are similar. So, PvuII's cutting site is located in H.1, namely, the longer part between two hincII sites.

$$\text{H.1} \underset{\text{PvuII}}{\rightarrow} \text{HP.1} + \text{HP.2}$$

Two Enzymes Conclusion We have got three conclusions:

1. EcoR1 is on shorter part between HincII. (Assumption with high likelihood)
2. EcoR1 is about the opposite of PvuII.
3. PvuII is on longer part between HincII.

In addition, it is mentioned above that EH.3/EH.4 are products of H.1 or H.2. And HP.1/HP.2 are products of H.1. However, both EH.3 and EH.4 is shorter than HP.1 and HP.2, for their smaller distances, which indicates that if we connect EH.3 with EH.4, the summation of length is less than HP.1 plus HP.2. That is, EH.3/EH.4 and HP.1/HP.2 come from two different reagents, EH.3/EH.4 from shorter one (H.2), and HP.1/HP.2 from longer one (H.1). So, here is the map.

Figure 11: Possible Map

2.3.3 Three Enzyme Digestion

For **EHP** tube, EHP.1 and EHP.2 are similar with HP.1 and HP.2. Meanwhile, EHP.3 and EHP.4 are similar with EH.3 and EH.4.

The bands correspond with the mapping result[3]:

$$\begin{aligned} \text{pBR322} &\underset{\text{HincII}}{\rightarrow} \quad \text{H.1} + \text{H.2} \\ \text{H.1} &\underset{\text{PvuII}}{\rightarrow} \quad \text{EHP.1} + \text{EHP.2} \\ \text{H.2} &\underset{\text{EcoR1}}{\rightarrow} \quad \text{EHP.3} + \text{EHP.4} \end{aligned}$$

2.4 Data Analysis

Before we start to calculate the length, we have to clean the raw length data (Table 3 and Table 4). As we have known that some bands have equal length, some are the summation of other bands, and some bands equal to the whole plasmid.

First, we will estimate the total length of plasmid by calculating the average of 7 summations.

[2] There is no order of two enzyme reactions, but explaining them in order can make reaction process clear.
[3] No fixed order in real reaction.

Table 5: Total Length of Each Well

Tube	E	H	P	EH	EP	HP	EHP
Length	3342	3112	3499	3041	2523	2089	2009
(bps)		1107		(1135)	2388	1552	1517
				621		1169	621
				430			423
total	3342	4219	3499	4092*	4911	4810	4570
average	4206						

* EH.2 (length = 1135) is remaining of incomplete reaction, not included.

So, the estimated total length is 4206bp. Assuming the total length is true, we can standardize all the lengths to their ratio to total length.

$$\text{Ratio} = \frac{\text{Length}}{\text{Total Length}} \in [0, 1]$$

Then the data will become:

Table 6: Ratio of All Fragments

Tube	E	H	P	EH	EP	HP	EHP
Ratio	1	0.738	1	0.743	0.514	0.434	0.44
		0.262		(0.27)	0.486	0.323	0.332
				0.152		0.243	0.136
				0.105			0.092

Then we will use a mathematical method solving equations to get the most probable result.

Since EHP.1/EHP.2/EHP.3/EHP.4 are 4 indivisible fragments and others are made from them, we assume their true ratio are x_1, x_2, x_3, x_4, where $x_1 + x_2 + x_3 + x_4 = 1$.

Thus, the assumption is,

Table 7: Ratio of All Fragments

Tube	E	H	P	EH	EP	HP	EHP
Ratio	1	$x_1 + x_2$	1	$x_1 + x_2$	- *	x_1	x_1
		$x_3 + x_4$		$(x_3 + x_4)$	- *	x_2	x_2
				x_3		$x_3 + x_4$	x_3
				x_4			x_4

* EP.1/EP.2's composition is unknown.

So, for all fragments, there are assumption ratio and experimental ratio. We can calculate the error by squared error:

$$\text{Error} = (\text{Ratio} - \text{Ratio}_{\text{Experi}})^2$$

Then, we can choose x_1, x_2, x_3, x_4 to minimize the error. The total error is calculated as:

$$\begin{aligned}\text{Error} =& (x_1 + x_2 - 0.738)^2 + (x_1 + x_2 - 0.743)^2 + (x_3 + x_4 - 0.262)^2 \\ &+ (x_3 + x_4 - 0.27)^2 + (x_3 + x_4 - 0.243)^2 + (x_1 - 0.434)^2 + (x_1 - 0.44)^2 \\ &+ (x_2 - 0.323)^2 + (x_2 - 0.332)^2 + (x_3 - 0.152)^2 + (x_3 - 0.136)^2 \\ &+ (x_4 - 0.105)^2 + (x_4 - 0.092)^2\end{aligned}$$

There are many methods to find the minimum of a function. I used the *optim()* function in **R** programming language to calculate the minimum error, which returned the result that:

$$\begin{cases} x_1 = 0.428 \\ x_2 = 0.319 \\ x_3 = 0.149 \\ x_4 = 0.104 \end{cases}$$

Now, we can know which fragments EP.1/EP.2 consist of.

$$\text{Ratio}_{\text{EP.1}} = 0.514 \approx x_1 + x_4 = 0.532$$

$$\text{Ratio}_{\text{EP.2}} = 0.486 \approx x_2 + x_3 = 0.468$$

That is, the assumption ratio of EP.1 and EP.2 should be $x_1 + x_4$ and $x_2 + x_3$.

After knowing that, we can do the optimization again, including the new ratio of EP.1 and EP.2. The result is:

$$\begin{cases} x_1 = 0.424 \\ x_2 = 0.323 \\ x_3 = 0.154 \\ x_4 = 0.099 \end{cases}$$

Then, using the optimized ratio, we can get the length by multiplying the Total length:

Table 8: Ratio of All Fragments

Tube	E	H	P	EH	EP	HP	EHP
Ratio	1	0.747	1	0.747	0.523	0.424	0.424
		0.253		(0.253)	0.477	0.323	0.323
				0.154		0.253	0.154
				0.099			0.099

Table 9: Length of All Fragments

Tube	E	H	P	EH	EP	HP	EHP
Length	4206	3142	4206	3142	2200	1783	1783
(bps)		1064		(1064)	2006	1359	1359
				648		1064	648
				416			416

Finally, we can get the mapping result Figure 4easily. Two HincII divide the plasmid into two parts, with length 3142 and 1064. EcoR1 is at the shorter part, dividing this short part into 648 and 416. PvuII is at the longer part, dividing this long part into 1783 and 1359. In addition, the 1783 and 416 is adjacent summing to 2200, and 1359 is adjacent summing to 2006. So the order is:

$$416 \to 1783 \to 1359 \to 648 \to ...$$

Figure 12: Mapping Result

2.5 Undigested DNA

There are two bands in **pBR322** tube, and the one with larger travel distance has stronger signal, the one with smaller travel distance has weaker signal.

The stronger signal one is the normal circular DNA, with supercoiled conformation, namely, more denser conformation. So it travels fast, regardless of its large number of base pairs. However, the weaker signal one is untwisted, less denser than supercoiled, causing it travel slower.

2.6 Conclusion

2.6.1 Common Error

Experimental Error As the electrophoresis does not exactly fit the equation $d \propto \frac{1}{log(n)}$, the distance itself consists of random error, which is inevitable in experiment. If we use wrong amount of reagents, or let reaction running for wrong time, or make mistakes when loading, the experimental error will increase.

Measurement Error There is no standard scale in the photo. I measure the distance using software to count the pixels, which will generate random measurement error. However, after measuring for enough times, calculating the average distance can eliminate the random measurement error.

Linear Fitting Error The standard curve is a curve, instead of a line. No matter how it looks like a straight line, the true value comes from a curve function. So if we use a linear function to estimate base pairs, there will always be error. Plus, the base pairs are in logarithm scale, which will magnify the small error into large length difference.

2.6.2 Further Application

At the end of the experiment, we get the locations of 4 restriction sites of 3 enzymes in a plasmid, and also, these fragments with known restriction ends. We can bind these fragments into other known plasmids to determine what genes are in the fragments. Then, we will get the gene map of the unknown plasmid. This is useful to study a new plasmid or test a newly built plasmid. In genetic engineering, we can use these plasmids with known genetic structure to do recombination with bacteria chromosomes and build new genotypes. This is the basic of transgenosis.

However, we have to admit that, compared with new generation of seqeuncing technology, the information we get from a traditional genetics experiment is not enough.