Chapter 3

Restriction Enzyme

"Will we ever be able to unravel these genes' labyrinthine workings? In his concluding chapter, Mr. Jacob asks whether there are limits to scientific research. For instance, it may slow down because it has become unmanageably large, "like a building that cannot rise to infinite heights". Or else there may be a limit to our understanding which could be "like a net that can catch only fish larger than its holes, or a microscope that cannot resolve details smaller than the resolving power of its lenses". There might indeed be a limit to the degree of complexity that we can comprehend, such as the interaction between thousands of genes or between the billions of nerves in our brains. Mr. Jacob fears that "the human brain may be incapable of understanding the human brain", because it is too complex. "

—The Economist **Genetic Science: Elusive Brains..., Dec 13 1997**

3.1 Restriction Enzyme

Type II sequence specific restriction endonucleases are enzymes that can "cut" a double-stranded DNA by breaking the phosphodiester bonds on the two DNA strands at specific target sites on the DNA. These target sites or "restriction sites" are determined completely by their base-pair composition—usually, a very short sequence of base-pairs with their lengths varying from 4 to 8. For instance, the restriction enzyme *Hpa* II will cut the DNA anywhere there is an occurrence of the tetranucleotide CCGG.

Note that if a site has been methylated even if the base pair composition corresponds to a restriction recognition subsequence the enzyme will fail to cut. Also when the enzymes cut at a site the two phosphodiester bonds may be in exact correspondence and the cleavage leaves two *blunt ends*; or, the two broken phosphodiester bonds may be slightly out of correspondence and the cleavage leaves exposed single stranded overhangs, called *sticky ends* as the hydrogen bonds on the singlestranded overhangs can make them to stick to complementary nucleotides.

The type II restriction endonucleases evolved in nature as the bacterial immune system against the viral DNA; bacteria use these enzymes by cleaving (or "restricting" the activity of) invading foreign DNA. Bacteria's own DNA is protected from its arsenal of restriction enzymes by another class of enzymes that modifies (or "methylates") the host DNA; as we know, restriction enzyme cleaves only unmethylated DNA.

The enzymes have been extremely useful in biotechnology as biochemical "scissors" and biochemical "markers" as they always cut DNA at the same short specific patterns (of length 4, 6 and 8). For instance, the first application of restriction enzymes in defining polymorphisms (sequence variations within a population) was by detecting the differences in length of small DNA fragments after digestion with sequence specific restriction endonucleases. The approach is called "restriction fragment length polymorphisms" (RFLP, pronounced aRFLiP), and is commonly used in forensic applications (e.g., from OJ to Clinton).

In wide use within the biotechnological laboratories, there are about 300 restriction enzymes, cutting at about 100 distinct restriction patterns. Note that most of these restriction patterns are of even length (> 2) and are "reverse palindromic"

$$s = \bar{s}^R$$

That is, the restriction patterns are invariant under reverse complementation. For instance, in the following example, the recognition sequence for *Hae* III is seen to have this reverse palindromy.

$$\overline{\text{GGCCGGCC}} = \text{CCGGCCGG}$$

and

$$\overline{\text{GGCCGGCC}}^R = (\text{CCGGCCGG})^R = \text{GGCCGGCC}.$$

If a restriction pattern is of length k, then the corresponding enzyme will be called a k-cutter; thus, a tetranucleotide-recognizing restriction enzyme is a 4-cutter; a hexanucleotide-recognizing restriction enzyme is a 6-cutter; and an octanucleotide-recognizing restriction enzyme is a 8-cutter.

Hae III:	GGCC|GGCC CCGG|CCGG	Blunt end 8-cutter
*Eco*R I:	G|AATTC CTTAA|G	Sticky end 6-cutter

The naming convention for the restriction enzyme is based on their occurrences in nature (rather unfortunate, since it leaves no clue as to the restriction patterns or the cutting frequency).

- The first three letters of a restriction enzyme: refer to the *microorganism* (e.g., *Eco* for E. coli, written italicized).
- The fourth letter: refers to the *strain* (e.g., R).
- Roman numerals index the restriction enzyme from the same organism (e.g., I).

Following is a more extensive table of restriction enzymes:

MicroOrganism	Restriction Enzyme	Restriction Site
Bacillus Amyloliquefaciens H	*Bam* HI	G|GATCC CCTAG|G
Brevibacterium albidum	*Bal* I	TGG|CCA ACC|GGT
Escherichia coli RY13	*Eco* RI	G|AATTC CTTAA|G
Haemophilus aegyptius	*Hae* II	P_uGCCG|P_y P_y|CGGCP_u
Haemophilus aegyptius	*Hae* III	GG|CC CC|GG
Haemophilus influenza Rd	*Hind* II	GTP_y|P_uAC CAP_u|P_yTG
Haemophilus influenza Rd	*Hind* III	A|AGCTT TTCGA|A
Haemophilus parainfluenza	*Hpa* I	GTT|AAC CAA|TTG
Haemophilus parainfluenza	*Hpa* II	C|CGG GGC|C
Providencia stuartii 164	*Pst* I	CTGCA|G G|ACGTC
Streptomyces albus G	*Sal* I	G|TCGAC CAGCT|G

One of the simplest things to try to do with restriction enzymes is to "map" a genome by marking the restriction sites on its DNA; this is also called the "restriction map". For instance, if we use a 6-cutter, we can expect to get several markers spaced about 4 Kb apart on the average. We may not be able to measure the restriction fragment lengths or equivalently, restriction cleavage locations exactly—this introduces the *sizing error* and determines the *accuracy* of the map. Sometimes, we may miss some percentage of the restrictions sites from the map or have them in the wrong order and may only be able to specify them in some partial ordering—this determines the *completeness* of the map. Finally, the spacing between the markers is directly related to how informative the map is and determines its *reso-*

lution. Both accuracy and completeness of a map can be easily improved by repeating the mapping experiments many times with many copies of the DNA molecules. The resolution of a restriction map can be improved by making maps with many different enzymes and combining the maps. For instance, one would expect that 4 different 6-cutter enzymes will lead to maps with a resolution of 1 Kb.

In general, the process of creating restriction maps or clone library with restriction enzyme can be affected by several error processes associated with cleavage:

1. *Partial Digestion*: In this case, the cleavage processes at all the restriction sites do not proceed to completion. This error process can be modeled as a Bernoulli trial, associating a probability $0 < p_c < 1$ that any given restriction site is cleaved. Thus, if a DNA has n restriction sites, the probability that we observe exactly k sites is given by the Binomial distribution

$$\binom{n}{k} p_c^k (1-p_c)^{n-k},$$

The moment generating function is given by

$$\begin{aligned}\Psi(t) &= \sum_{k=0}^{n} \binom{n}{k} (p_c e^t)^k (1-p_c)^{n-k} \\ &= \left[1 + p_c(e^t - 1)\right]^n\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}[k] &= \Psi'(0) \\ &= n\left[1 + p_c(e^t - 1)\right]^{n-1} p_c e^t|_{t=0} \\ &= np_c \\ \mathbb{E}[k^2] &= \Psi''(0) \\ &= n(n-1)\left[1 + p_c(e^t - 1)\right]^{n-2} p_c^2 e^{2t}|_{t=0} \\ &\quad + n\left[1 + p_c(e^t - 1)\right]^{n-1} p_c e^t|_{t=0}\end{aligned}$$

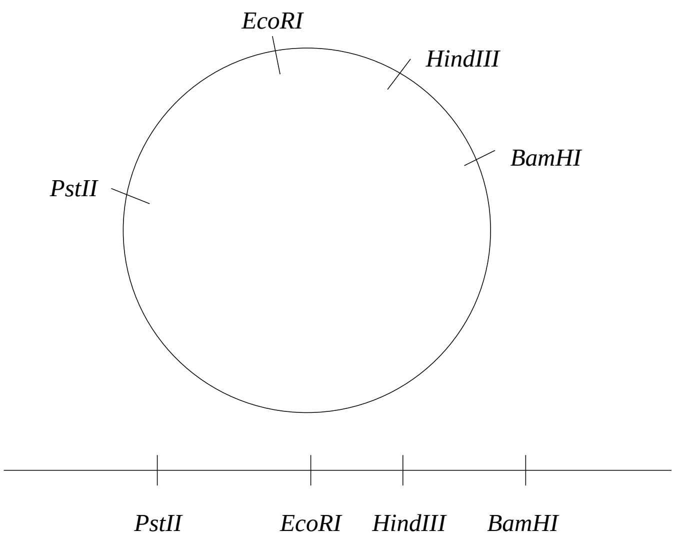
$$
\begin{aligned}
&= n^2p_c^2 - np_c^2 + np_c \\
\text{Var}[k] &= np_c(1-p_c) \\
\text{Std. Dev.}[M] &= \sqrt{np_c(1-p_c)}
\end{aligned}
$$

2. *False Cuts & Star Activity*: In some cases, one may get a mechanical breakage at a site not corresponding to a recognition pattern. Or equivalently, the sensors involved in finding the cleavage site may mistakenly label a non-recognition site as a restriction site; this is quite common in many single-molecule approaches. The error process involved in this can be modeled as a Poisson process.

   Star activity on the other hand corresponds to the loss of restriction enzyme specificity—particularly, when the enzyme is used under altered chemical conditions: such as, high enzyme concentration, high pH or the presence of organic solvents like glycerol. The main effect of star activity is to cleave DNA at sequences that differ from the normal recognition sequence by one or two base pairs. For instance, a given 6-cutter may also become specific to the tetranucleotide subsequence at the center of its normal hexanucleotide recognition sequence. This can be easily modeled as independent Bernoulli trials.

3. *Missing Fragments*: Another problem that makes collection of a complete set of restriction fragments elusive is that some of the small fragments escape detection, or fail to be part of the collection prepared for subsequent chemical processes etc. Often the end fragments pose significantly more difficulties than the others.

A map can be presented as in the following diagram, showing a restriction map of a the *Plasmid* `pBR322` in E. coli. This is a circular DNA of length 4363bp. Thus one would expect the 4 6-cutters (*Eco* RI, *Hind* III, *Pst* II and *Bam* HI) to cut this DNA in about one place each. The resolution of this map is about 1 Kb.



3.2 Restriction Map Resolution

3.2.1 Number of Restriction Sites

We start by looking at the statistical distribution of the number of restriction sites in a genomic DNA of length G. By p_k, we denote the *cutting frequency*, and is the probability that an arbitrary site (i th position in the genomic DNA, $1 \leq i \leq G - k$) is a restriction site (or more precisely the 5' end of the site) for the k-cutter restriction enzyme.

Assuming that all base pairs occur at any given position with equal probability and independently (uniform i.i.d.), we see that with the restriction pattern $s \in (\mathrm{A} + \mathrm{T} + \mathrm{C} + \mathrm{G})^+$:

$$p_k = \frac{1}{4^k}, \quad k = |s| \in \{4, 6, 8\}.$$

Numerically:

$$p_4 = \frac{1}{256}, \quad p_6 = \frac{1}{4,096}, \quad \text{and} \quad p_8 = \frac{1}{65,536}.$$

Let $\lambda_k = Gp_k$, where G is the length of a genomic DNA to be cut by a restriction enzyme. Thus the number of restriction sites in the DNA has a *Binomial distribution*.

$$\Pr[\exists \ M \text{ restriction sites in a DNA of size } G] = \binom{G}{M} p_k^M (1-p_k)^{G-M}.$$

A somewhat simpler Poisson approximation would be given by the following application of *Brun's Sieve*.

Theorem 3.2.1 Brun's Sieve: *Let W be a nonnegative integer-valued random variable such that*

$$\mathbb{E}\left[\binom{W}{i}\right] \approx \frac{\lambda^i}{i!}$$

Then

$$\Pr[W = M] \approx e^{-\lambda}\frac{\lambda^M}{M!}. \qquad \square$$

Let X_i $(1 \leq i \leq G)$ be a Bernoulli random variable denoting the event that "there is a restriction site beginning at i" ($X_i = 1$; otherwise, $X_i = 0$). Then

$$W = \sum_{i=1}^{G} X_i,$$

denotes the total number of restriction sites in the genome.

$$\Pr[X_i = 1] = p_{i,k} = 1 - \Pr[X_i = 0].$$

We assume that $p_{i,k} \approx p_k \ll 1$ and are "very weakly" dependent: $p_{i_1,k} p_{i_2,k} \cdots p_{i_l,k} \approx p_k^l$.

Interpret W as the number of successes in G independent trials, where each is a success with probability p_k. Then $\binom{W}{i}$ is the number of i successful trials. Now consider the set of all i trials—there are $\binom{G}{i}$ of these, and each i successful trials occur with probability $\approx p_k^i$.

$$\mathbb{E}\left[\binom{W}{i}\right] \approx \binom{G}{i} p_k^i \approx \frac{G^{(i)}}{i!} p_k^i \approx \frac{(Gp_k)^i}{i!}.$$

Using Brun's sieve, we have:

$$\begin{aligned}\Pr[\#\text{restriction sites} = M | G, \lambda_k] &= Prob[W = M] \\ &\approx e^{-\lambda_k}\frac{\lambda_k^M}{M!}.\end{aligned}$$

Proof of Brun's Sieve:.

Let $\mathbb{I}_{W=j}$ be

$$\mathbb{I}_{W=j} = \begin{cases} 1 & \text{if } W = j; \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$\begin{aligned}\mathbb{I}_{W=j} &= \binom{W}{j}\sum_{k=0}^{W-j}\binom{W-j}{k}(-1)^k \\ &= \sum_{k=0}^{W-j}\binom{W}{j}\binom{W-j}{k}(-1)^k \\ &= \sum_{k=0}^{\infty}\binom{W}{j+k}\binom{j+k}{k}(-1)^k\end{aligned}$$

By convention $\binom{W}{j} = 0$, if $j > W$.

Thus,

$$\begin{aligned}\Pr[W = j] = \mathbb{E}\left[\mathbb{I}_{W=j}\right] &= \sum_{k=0}^{\infty}\mathbb{E}\left[\binom{W}{j+k}\right]\binom{j+k}{k}(-1)^k \\ &= \sum_{k=0}^{\infty}\frac{\lambda^{j+k}}{(j+k)!}\binom{j+k}{k}(-1)^k \\ &= \frac{\lambda^j}{j!}\sum_{k=0}^{\infty}\frac{-\lambda^k}{k!} \\ &= e^{-\lambda}\frac{\lambda^j}{j!}.\end{aligned}$$

$\square$

Thus we have the following statistics for the number of restriction sites in genomic DNA of length G. The moment generating function for the Poisson distribution is given by

$$\begin{aligned}
\Psi(t) &= \mathbb{E}[e^{tM}] \\
&= \sum_{M=0}^{\infty} e^{tM} e^{-\lambda_k} \frac{\lambda_k^M}{M!} \\
&= \sum_{M=0}^{\infty} e^{-\lambda_k} \frac{(\lambda_k e^t)^M}{M!} e^{-\lambda_k e^t} e^{\lambda_k e^t} \\
&= e^{\lambda_k(e^t-1)} \sum_{M=0}^{\infty} e^{-\lambda_k e^t} \frac{(\lambda_k e^t)^M}{M!} \\
&= e^{\lambda_k(e^t-1)}
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}[M] &= \Psi'(0) \\
&= \lambda_k e^t (e^{\lambda_k(e^t-1)})|_{t=0} \\
&= \lambda_k = Gp_k = G/4^k \\
\mathbb{E}[M^2] &= \Psi''(0) \\
&= \left(\lambda_k e^t (e^{\lambda_k(e^t-1)}) + \lambda_k^2 e^{2t} (e^{\lambda_k(e^t-1)})\right)\Bigg|_{t=0} \\
&= \lambda_k + \lambda_k^2 \\
\mathrm{Var}[M] &= \lambda_k^2 + \lambda_k - \lambda_k^2 = \lambda_k \\
\text{Std. Dev.}[M] &= \sqrt{G}/2^k.
\end{aligned}$$

3.2.2 Size of Restriction Fragments

The number of base pairs in a restriction fragments can be determined as a follows: Start with a restriction site, corresponding to the left end (5' end) of a restriction fragment, and proceed towards the 3' end while counting the base pairs. The probability that no other restriction site is encountered over the next l base pairs and that there is a restriction site starting at the $(l+1)$th base pair is given by

$$(1-p_k)^l p_k \approx \frac{e^{-l/\mu_k}}{\mu_k}$$

where $\mu_k^{-1} = \log(1/(1-p_k))$.

In particular, let W be a random variable with exponential distribution with mean μ_k and $f_W(w) = \frac{e^{-w/\mu_k}}{\mu_k}$, $w > 0$, then $Z = \lfloor W \rfloor$ stands for the random variable giving the length of a restriction fragment in base pairs.

Thus we have the following statistics for the length of a restriction fragment. The moment generating function for the exponential distribution is given by

$$\begin{aligned}
\Psi(t) &= \mathbb{E}[e^{tw}] \\
&= \int_0^\infty e^{tw} \frac{e^{-w/\mu_k}}{\mu_k}\, dw \\
&= \frac{1}{1-\mu_k t} \int_0^\infty e^{((1-t\mu_k)/\mu_k)w} \left(\frac{1-t\mu_k}{\mu_k}\right) dw \\
&= \frac{1}{1-\mu_k t}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}[W] &= \Psi'(0) \\
&= \left(\frac{\mu_k}{(1-\mu_k t)^2}\right)\Bigg|_{t=0} \\
&= \mu_k \approx \frac{1}{p_k} = 4^k \\
\mathbb{E}[W^2] &= \Psi''(0) \\
&= \left(\frac{2\mu_k^2}{(1-\mu_k t)^3}\right)\Bigg|_{t=0} \\
&= 2\mu_k^2 \\
\mathrm{Var}[W] &= 2\mu_k^2 - \mu_k^2 = \mu_k^2 \\
\text{Std. Dev.}[W] &= \mu_k \approx 4^k.
\end{aligned}$$

3.2.3 Matching Rules for Restriction Fragments

Given two restriction fragments without any other identifying markers, when can we say that they are the same? A plausible rule would be to simply compare the lengths and declare that they are the same if they have exactly the same lengths. In the presence of measurement errors, of course, this would make the probability that we have a "false negative" match arbitrarily close to 1. That is, we will almost surely fail to match two identical restriction fragments whose lengths have been measured with some error. Thus, we must account for the measurement errors and compare the lengths modulo some small "relative sizing error parameter" β. The following simple rule is frequently used:

Matching Rule (II). Two restriction fragments are said to match if their lengths x and y differ by less than 100 β%.

$$-\beta \leq 1 - \frac{y}{x} \leq \beta.$$

That is, $y \in [x - \beta x, x + \beta x]$. Equivalently, $x \in [\frac{y}{1+\beta}, \frac{y}{1-\beta}]$. For small β, we see that the matching rule II is "symmetric." □

Thus if we choose sufficiently large β, then the "false negative" match probability gets arbitrarily close to zero: That is if the restriction fragments are indeed the same, the matching rule would then match them almost surely. By $\delta_\beta \approx 0$, we deonote this false negative probability.

However, we now need to quantify the "false positive" match probability. In other words, given two randomly chosen distinct restriction fragments obtained by cleaving a "large" genomic DNA by the same restriction enzyme, what is the probability that the matching rule accidentally identify them as the same?

Thus we have the following:

- Expected length of a restriction fragment $= \mu_k$. Hence, both x, $y \sim$ Exponential$(1/\mu_k)$.

$$f_X(x) = \frac{e^{-x/\mu_k}}{\mu_k}.$$

- Now the probability that two randomly chosen restriction fragments match (up to $100\beta\%$) by the matching rule II, is given by:

$$
\begin{aligned}
&= \int_0^\infty \left(\int_{x(1-\beta)}^{x(1+\beta)} \frac{e^{-y/\mu_k}}{\mu_k}\, dy \right) \frac{e^{-x/\mu_k}}{\mu_k}\, dx \\
&= \int_0^\infty \left(\int_{v(1-\beta)}^{v(1+\beta)} e^{-w} dw \right) e^{-v} dv \\
&= \int_0^\infty \left[-e^{-v(1+\beta)} + e^{-v(1-\beta)} \right] e^{-v} dv \\
&= -\frac{1}{2+\beta} \int_0^\infty e^{-v(2+\beta)}(2+\beta)\, dv \\
&\quad + \frac{1}{2-\beta} \int_0^\infty e^{-v(2-\beta)}(2-\beta)\, dv \\
&= \frac{1}{2-\beta} - \frac{1}{2+\beta} \\
&= \frac{2\beta}{4-\beta^2} \approx \frac{\beta}{2}.
\end{aligned}
$$

In order to get a realistic estimate, imagine that you are conducting an experiment with a six cutter restriction enzyme, making the value of $\mu_k = 4^6 = 4,096 \approx 4Kb$. A realistic sizing error would be about $700bp$ and we may wish to apply our matching rule with a $\beta = 1/4$. Then the above calculation would show that with probably 1/8 you will get a false positive match.

3.2.4 Reality Check

At this point, it is worth mentioning that our assumption that all four base pairs have equiprobable and independent occurrences, is often false. This is to be expected as the genome in a DNA encode for various functionality and exhibit many structural motifs at many different scales. We will come back to the structure of the genome in another lecture.

In reality, the distribution of nucleotides is non-random, in that certain base pairs (say, G + C—GC content) occur significantly more or significantly less frequently than the other (A+T).

For instance, in human G+C occur with probability of about 0.37 and the dinucleotide CpG occur with a probability about 1/5th of the expected value. Consequently, the 8-cutter restriction enzyme *Not* I, with the recognition pattern GCGGCCGC, when applied to human DNA yields fragments of lengths 1–1.5 Mb, a range significantly higher than the expected value of $4^8 \approx 65Kb$. Similarly, E. coli genome of size $G = 4.7 \times 10^6 bp$ is cut by *Not* I into only 20 restriction fragments and not the expected number

$$\frac{G}{p_k} = \frac{4700}{65} \approx 70.$$

For this reason, *Not* I is labelled a *rare cutter* and finds a particularly significant role in many biotechnological applications.

Another simple discrepancy is accounted for by the fact that in G + C rich genomes, the tetranucleotide CTAG is rare and many restriction enzymes containing this tetranucleotide in their recognition patterns cut less frequently than the p_k value. Examples of such enzymes are:

Restriction Enzyme	Restriction Site
Bln I	CCTAGG
Nhe I	GCTAGC
Spe I	ACTAGT
Xba I	TCTAGA

Other interesting examples come from the considerations of the intron-encoded endonucleases, such as I-*Cen* I or I-*Sce* I, which have very long recognition patterns ranging between 18–30 bp and many of its cleavage sites are within related genes. For instance, I-*Cen* I, from the chloroplast large rRNA gene, cuts within the *rrn* operons. Additionally, these restriction enzymes, by their nature, are extremely rare cutters.

Another useful technique to modulate cutting frequency, is the so-called RARE (RecA-assisted restriction endonuclease) technique. In this approach, a particular locus on the DNA containing the restriction site of the given restriction enzyme is inactivated by binding the RecA protein to the chosen locus.

The DNA is then methylated by the methyltransferase and the RecA protein is removed. Now, only the restriction sites within the chosen locus is exposed for cleavage.

Of course, a simple approach to reduce the number of cutting sites is by not letting the digestion process to last until completion. However, the *partially digested* DNA cannot be very precisely controlled and its potential usages are in library formation and map making.