# Chapter 4

# Maps

> *"Trying to understand the way nature works involves a most terrible test of human reasoning ability. It involves subtle trickery, beautiful tightropes of logic on which one has to walk in order not to make a mistake in predicting what will happen."*
>
> —Richard Feynman, *The Meaning of It All*, pp. 15.

## 4.1 Finger Prints and Maps

We first introduce the concept of finger prints and maps of a clone or a genome. For our purpose here, a clone is a rather large fragment of the DNA that have been pre-selected and kept in a library, and one can make faithful copies of this DNA fragment many many times. The size of a clone can be 1–2 Mb (YAC), 100–200 Kb (BAC), 20–45 Kb (Cosmids) or 2–20 Kb (lambdas).

| Vector | Insert Size |
|---|---|
| Lambda | 2–20 Kb |
| Cosmid | 20–45 Kb |
| BAC (Bacterial Artificial Chromosome) | 100–200 Kb |
| BAC (Yeast Artificial Chromosome) | 1–2 Mb |

If we take one such clone and completely digest it into small pieces by a restriction enzyme, then since the enzyme cuts only at fixed sites, the set of restriction fragments and their order is always the same for that clone. The unordered collection of these restriction fragments is called a *finger print* or an *unordered restriction map* and the ordered set is called an *ordered restriction map*.

## 4.2 Cloning

We briefly discuss how the clones are copied and how the clone finger prints or maps are obtained. This will lead us to a rather simple but an important question: given two clones, when can we say whether they overlap by simply looking at their finger prints or maps (which may be subject to sizing errors).

### 4.2.1 Amplification by Molecular Cloning

The molecular cloning is an *in vivo* approach involving a living host organism (usually the E. coli bacteria or yeast) which replicates a suitably modified foreign DNA, as if the foreign DNA is one of its own DNA. The modification involves combining a cloning "*vector*" with the foreign DNA to be amplified, the "*insert*," to create a circular recombinant DNA molecule, the "*replicon*;"—the cell will not replicate any foreign DNA in the absence of a suitable vector.

The key to this approach involving the lambda clones is the process of transduction: in transduction, the DNA from the genome of one strain of E. coli would get incorporated into the genome of a phage and the modified phage would then get replicated in a different strain, thus making it (e.g., the lambda phage) an ideal vector. Cosmid clones are, on the other hand, artificially created circular DNA that acts like a plasmid (much smaller extrachromosmal DNA found and replicated in bacteria; they can appear either as single copy and get replicated as the host organism replicates or as multiple copies and replicate independently). BAC's and YAC's are artificially created DNA that act more like chromosomes in the host organism—BAC's (Bacterial Artificial Chromosomes) are replicated in bacteria such as E. coli; YAC's (Yeast Artificial Chromosomes) are replicated in the single-cell eukaryotic S. cerevisae (baker's yeast).

Here are some of the details of the steps in molecular cloning:

1. *Step 1, Recombinant DNA Replicon*: Linear inserts and the short (also linearized) vectors with same sticky ends are mixed together along with DNA ligase enzyme. The sticky ends of the inserts attach to the sticky ends of the vector as a result of the hydrogen bonding and is ready to be replicated. Of course, some of the inserts stick to its own ends (thus *recyclizing*), but fortunately do not get replicated by the host. Consider plasmid vectors—for instance, plasmids containing a single *Eco* RI restriction site embedded within the gene for ampicillin-resistance. Digestion with *Eco* RI produces linearized plasmids with stick ends $|$AATT $\cdots$. Now consider inserts of length about 4 Kb obtained by digestion with *Eco* RI, which also have the same sticky ends: $|$AATT $\cdots$. When two sticky ends meet, $|$AATT $\cdots$ with $\cdots$ TTAA$|$ (preferably between an end of the vector with an end of the insert), they will bond making the replicon.

2. *Step 2, Transforming and Growing the Cells*: Next step is to transform the host cell by inserting the replicon. In case of the phage, the process is simple, efficient and well-understood. But it is still unknown how the plasmids or the BACs and YACs enter a cell, but number of well-reproducible protocols have been developed that involve exposing a population of the host cells to the ligase mixture containing the replicons. Note that in many cases (e.g., single copy plasmids), it is necessary that only one copy of the replicon enters the cell.

   A small amount of the diluted solution containing the transformed cells is transferred to culture dishes containing solid growth medium, and the cells divide making a colony of the original transformed single cells. The cells in the colony each contains at least on copy of the replicon and thus the insert. The host organisms normally used, E. coli and S. cerevisae, have short *generation time* of about 20 minutes and can yield as much as a billion inserts ($2^{30}$) in 10 hours.

3. *Step 3, Colony Picking*: Next job is to identify the copies (or clones) of the transformed cells containing the copies of the inserts. For instance, if we are using the plasmid vectors with an *Eco* RI site in its ampicillin resistance gene, then the transformed cells have plasmids whose ampicillin resistance genes have been interrupted and made inoperative, thus making them susceptible to death by ampicillin attack. This fact can be used to identify regions of the culture dish containing the copies of the transformed cells, for instance, by sampling the culture dish, often with specialized robots.

   Finally, the cells are identified and colonies, picked; the replicons are isolated and linearized (say by cutting with *Eco* RI enzyme, as the vector-insert boundaries contain *Eco* RI recognition sequences).

## 4.3 Gel-Electrophoresis and Finger Printing

Next we describe a rather simple approach to see how one can finger print a clone (i.e., collect the set of restriction fragments of the clone without any order information). The main idea is to take a collection of clones of a particular insert in a test tube and treat them with a restriction enzyme under the ideal chemical conditions. If the clones are all completely digested, then we now have a large number of the same restriction fragments in the test tube and it remains to simply distinguish them. Of course, the digestion process in the test tube destroys any order information (even whether two fragments are consecutive, since we assume that the digestion is complete).

Note that the phosphate groups of a DNA fragment are negatively charged and under a constant electric field will experience a force proportional to its length and hence to its mass and in the direction opposite to the field. However, in a free environment, this would make all the fragments move with the same acceleration and would not help much in distinguishing the restriction fragments by mass. But if the fragments are made to move through a porous semi-solid *gel* medium under a constant electric field, then the fragments experience an additional damping force proportional to the logarithm of its length. Thus, to a rough approximation, the drift rate of a population of restriction fragments of equal length under the constant electric field and through the gel is inversely proportional to the logarithm of its length. Often an approximate formula used to estimate how far a fragment of length $L$ would move in gel-electrophoresis is given as

$$D(L) = a - b \log L, \quad a, b > 0$$

thus making it clear that the small fragments will move the farthest as they will not entangle with the gel matrix much and the large fragments will move the least in the matrix.

—`Needs a clearer analysis.`

Depending on the average lengths of the fragments to be separated (e.g., in the range of few hundred bps or few tens of Kbs), different gels need to be used—for instance, agarose gel (solidified aqueous solution of purified agar) is used for large fragments, and polyacrylamide gel for small fragments as needed in sequencing.

The positions of the fragments is detected by soaking the gel in a solution of ethidium bromide, which bonds with DNA and emits visible light when excited by UV light. The diffused volumes of fragments of DNA of same length (and not necessarily same sequence) appear as light bands in the gel images. By comparing the positions of the bands with the set of bands in a lane containing gel-calibration sample of fragments of known length and using the expression for $D(L)$ given earlier, one computes a finger print.

In a widely-used variation, where long fragments (more than

50 Kb) are involved, the electric field is pulsed between two fixed values (0 and some fixed value, say) and also varied directionally locally by small magnitude. This has the effect of reducing the diffusion rate for the large fragments, while improving the drift and thus allowing the large fragments to be separated into detectable bands. The variant method is called "PFGE" (Pulsed-Field Gel-Electrophoresis).

## 4.4 Matching Clones

We are now interested in the following simple question. Suppose we have two clones: Clone $A$ and clone $B$ of (roughly) the same length obtained from a genome by some process that does not preserve any location information. But there is a possibility that they do overlap; perhaps they were obtained by mechanical shearing or by partial digestion—note that if $A$ and $B$ were created by complete digestion, then we already know that they do not overlap. A simple process would involve making an ordered restriction map or a finger print for each of the clones $A$ and $B$ and then check for an event that is highly correlated with overlap event. For instance, with two maps we could compare the ordered restriction fragments at the ends or with the fingerprints we could take the set intersection of the fingerprints and see if it has a large cardinality. Of course, we need to accomodate for the sizing error, measured in terms of the parameter $\beta$: *the relative sizing error*.

We first define an "*overlap threshold ratio*," $\theta = T/L$, where $L$ is the length of the clones (all equal) and $T$ is the "overlap length parameter" that determines how many base pairs the two clones must have in common to ensure their overlap. Thus for clones $A$ and $B$, we say they "overlap" if and only if

$$\frac{|A \cap B|}{(|A| + |B|)/2} \geq \theta.$$

Thus a "false negative" only means that the two clones do overlap in this restricted sense and yet we fail to recognize that.

Let $n$ denote the number of restriction fragments in each clone. Thus the number of fragments in a subregion of length $\theta L$ is distributed as a Binomial distribution $\sim S(n, \theta)$ and if we choose $n' = n\theta/2$ then, by the Chernoff bounds, the probability that $n'$ or more fragments occur in the subregion is greater than

$$\Pr[S(n, \theta) \geq n'] \geq 1 - e^{(-n\theta/8)}.$$

### 4.4.1 Comparing Finger prints

As stated earlier, we decide that two clones $A$ and $B$ overlap if they have at least $k$ (where $k \leq \lfloor n\theta/2 \rfloor$) restriction fragments in common. It is rather easy to see that if in fact the clones "overlap" then we succeed with probability close to 1.

Now the next question we need to answer is when two randomly chosen unrelated clones may get determined to overlap.

Let $W$ denote the number of restriction fragments in clone $A$ with the property that each one of these restriction fragments find a match with a *distinct* fragment of clone $B$. Note that the random variable $\binom{W}{i}$ represents the number of $i$ fragments that all match to the fragments in the other clone $B$.

For each of the $\binom{n}{i}$ sets of "clone-$A$-$i$-fragments" define an indicator variable that evaluates to 1 if all the fragments of this set find matches and 0, otherwise.

$$X_j, \quad j = 1, \ldots, \binom{n}{i},$$

is the indicator variable for the $j$th set of "clone-$A$-$i$-fragments."

$$\binom{W}{i} = \sum_j X_j,$$

and

$$\begin{aligned}
\mathbb{E}\left[\binom{W}{i}\right] &= \binom{n}{i}\left[n\left(\frac{\beta}{2}\right)\right]\left[(n-1)\left(\frac{\beta}{2}\right)\right]\cdots\left[(n-i+1)\left(\frac{\beta}{2}\right)\right] \\
&= \frac{(n^{(i)})^2}{i!}\left(\frac{\beta}{2}\right)^i
\end{aligned}$$

$$= \frac{(\beta n^2/2)^i}{i!}$$

Now by another application of Brun's sieve, we have

$$\Pr[W = i] = \frac{1}{i!}(\beta n^2/2)^i e^{-\beta n^2/2}$$

a Poisson Distribution with a parameter $\beta n^2/2$. Thus in order to keep the tail probability representing the false positive overlap match small, we need to make $k$ at least $\lceil(3\beta n^2/4)\rceil$.

Note that, we need the following conditions satisfied in order for the fingerprints to effectively work in determining overlaps:

$$\frac{3\beta n^2}{4} \le k \le \frac{n\theta}{2}.$$

Thus

$$1 \ge \theta \ge \frac{3\beta n}{2} \quad \text{and} \quad \beta \le \frac{2}{3n}.$$

Thus fingerprints work only when the number of fragments is rather small and for large clones, this implies that very "rare" cutters are used. Furtherore, the "overlap threshold parameter" $\theta$ must be rather large. This has some serious negative implications, that we will see momentarily.

### 4.4.2 Comparing Restriction Maps

Just as before, we decide that two clones $A$ and $B$ overlap if they have at least $k$ (where $k \le \lfloor n\theta/2 \rfloor$) restriction fragments in common and they appear in the correct order in one end or the other. It is again easy to see that if in fact the clones "overlap" then we succeed with probability close to 1.

But, when we compare two unrelated randomly chosen unrelated clones they have exactly $i$ fragments matching in order in one end or the other is exactly

$$4(\beta/2)^i,$$

as there are 4 different relative orientations to consider.

Thus the probability that a declared overlap is a false positive is given by

$$\begin{aligned}&\sum_{i=k}^{\infty} 4(\beta/2)^i \\ &= \frac{4(\beta/2)^k}{1-\beta/2} \approx 4(\beta/2)^k[1+\beta/2]\end{aligned}$$

If the false positive probability to be kept below some small error probability $\epsilon$ then $k$ should be approximately $(\ln \frac{4}{\epsilon})(1-\beta/2)$.

Note that, we need the following conditions satisfied in order for the maps to effectively work in determining overlaps:

$$\left(\ln \frac{4}{\epsilon}\right)(1-\beta/2) \leq k \leq \frac{n\theta}{2}.$$

Thus

$$1 \geq \theta \geq \frac{2}{n} \ln \frac{4}{\epsilon}(1-\beta/2).$$

Thus maps work quite well for the larger clones and less sensitive to the sizing errors. This makes restriction maps a very powerful tool.

### 4.4.3 Comparing Partially Digested Restriction Maps

Now consider a fairly simple situation where we have restriction maps with some number of the restriction sites going undetected—this is the situation, if for instance the molecules are only partially digested. In a pioneering experiment, we attempted and succeeded to map the genomes of many micro-organisms by taking genomic DNA molecules and creating rough maps with high $\beta$ and low partial digestion $p_c$.

Thus we need to model the maps by several parameters: relative sizing error, $\beta$, partial digestion probability, $p_c < 1$, number of restriction fragments per clone, $n$ and number of detected restriction fragments, $m = np_c$. The overlap threshold ratio is $\theta$.

We need to choose an overlap rule that takes the partial digestion into account. The rule is fairly simple and as follows: we postulate an alignment, where clone $A$ is aligned with respect to clone $B$ with an overlap ratio bigger than $\theta$ (in 4 possible relative orientations); at least $k$ of the restriction fragments of the clones match positionally and the numbers of unmatched fragments in the prefixes are bounded by $r$.

—`See Figure.`

By an arguments similar to the ones earlier, we see that

$$k \leq \frac{np_c^4\theta}{2} \quad \text{and} \quad r \geq \frac{k_1}{p_c^4}, \quad k_1 \approx 2.$$

If in fact clones $A$ and $B$ overlap, then we will detect it with probability bigger than

$$(1 - e^{-k_1})(1 - e^{-np_c^4\theta/8}).$$

Now consider the false positive probability and consider an arbitrary alignment (not necessarily satisfying the constraints on the unmatched prefixes). Let the random variable $W$ denote the number of fragments in clone $A$ that positionally match with the fragments of clone $B$. Proceeding with an argument similar to the ones used earlier, we see that

$$\begin{aligned} \mathbb{E}\left[\binom{W}{i}\right] &= \binom{m}{i}(\beta/2)^i \\ &= \frac{1}{i!}\left(\frac{np_c\beta}{2}\right)^i. \end{aligned}$$

By the use of the Brun's sieve, we see that

$$\Pr[W = i] = \frac{1}{i!}(\beta np_c/2)^i e^{-\beta np_c/2}.$$

Thus the false positive probability is

$$4r\sum_{i=k}^{\infty}\frac{1}{i!}(\beta np_c/2)^i e^{-\beta np_c/2}.$$

and we need to make $r$ as small as possible and $k$ as large as possible.

Thus we need to satisfy the following constraints:

$$\frac{3\beta n p_c}{4} \le k \le \frac{n p_c^4 \theta}{2} \quad \text{and} \quad \frac{k_1}{p_c^4} \le r \le \frac{1}{\beta}.$$

Thus

$$1 \ge \theta \ge \frac{3\beta}{2p_c^3} \quad \text{and} \quad \beta \le \frac{2p_c^3}{3}.$$

Thus partially digested restriction maps work well only when the partial digestion probability is rather high (close to 1), i.e.,

$$p_c \ge (3\beta/2)^{\frac{1}{3}}.$$

or the relative sizing error is quite low, which can be achieved by making the fragment length larger by using a rare cutter.

## 4.5 Covering the Genome with Clones

Consider a genome of length $G$ that has been uniformly randomly sampled to collect $N$ clones each one of length $L$. The parameters of interest are summarized as follows:

$G =$ Genome length (in bp).

$L =$ Length of a clone.

$N =$ Number of clones.

$c = \left(\frac{LN}{G}\right) =$ *Coverage*.

Let the indicator variable $X_{i,j}$ denote the event that the clone $i$ covers the position $j$ of the genome:

$$X_{i,j} = \begin{cases} 1 & \text{if clone } i \text{ covers the base pair } j \\ 0 & \text{otherwise} \end{cases}$$

Let $W_j = \sum X_{i,j}$ be the random variable denoting the number of clones covering the position $j$. Thus

$$\mathbb{E}\left[\binom{W_j}{n}\right] = \binom{N}{n}\left(\frac{L}{G}\right)^n \approx \left(\frac{NL}{G}\right)^n /n! = \frac{c^n}{n!}.$$

Hence, by Brun's sieve, we have:

$$\Pr[W_j = k] = e^{-c}\frac{c^k}{k!}.$$

Thus the expected fraction of the genome that is represented in the clones is

$$f = \frac{\sum_{j=1}^{G} \Pr[W_j \neq 0]}{G} = 1 - \Pr[W_j = 0] = 1 - e^{-c}.$$