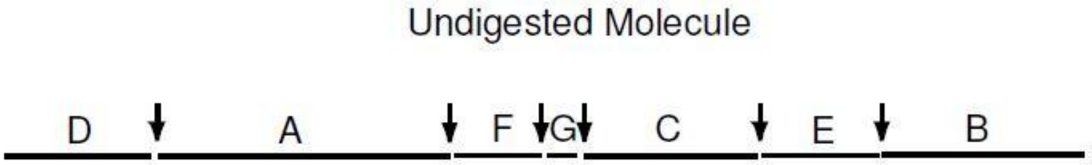
Chapter 4 Class Notes – Physical Mapping of DNA

4.1. The Biological Problem: physical versus genetic maps:

- **physical map: use physical distances (bp or kb); physical marker (e.g. restriction enzymes) allow researchers to retrieve particular regions of interest for study**
- **genetic map: genetic marker defined by mutations in genes; the presence of such markers can be detected by hybridization* reactions even if complete DNA sequence is unknown (* process to find DNA piece by using a probe)**

Chapter 3: cut DNA using restriction enzymes; here, we're interested in reconstructing the restriction map as below:



4.2. The Double-Digest Problem (DDP): a sample of DNA is digested with a restriction enzyme until all sites on every molecule are cut. Agarose gel electrophoresis gives a way to separate DNA fragments by their lengths, with small pieces migrating through the porous agarose quickly (distance proportional to the negative log of the lengths), and larger pieces hardly moving at all. So, gel electrophoresis gives us the set of lengths unordered relative to actual location in the DNA. E.g., digesting phage lambda DNA (48,502bp) with *Eco*RI (with recognition sequence 5'-G/AATTC-3') gives the lengths in kb: 3.5, 4.9, 5.6, 5.8, 7.4 & 21.2; correct order is 21.2-4.9-5.6-7.4-5.8-3.5.

One way to determine this ordering is via double-digests: using two restriction enzymes (denoted A and B) separately and in combination (A&B); possible orderings cans then be inferred:

<u>Example I</u>:

products produced from digestion with A alone: {2,4}
products produced from digestion with B alone: {1,5}
products produced from digestion with A&B: (1,1,4}

Note first that due to the ‘4’ in A&B, the ‘4’ and ‘5’ must lie on the same end of A and B. So, the only possible choice is below left (and its biologically indistinguishable reversal, below right):

A: 2 4	A: 4 2
B: 1 5	B: 5 1

<u>Example II</u>:

products produced from digestion with A alone: {3,4,5}
products produced from digestion with B alone: {2,4,6}
products produced from digestion with A&B: (1,1,2,3,5}

Again, the ‘5’ and ‘6’ must lie on the end (chosen below to be the right end). This gives the following 4 possibilities:

A:	3 4 5	3 4 5	4 3 5	4 3 5
B:	2 4 6	4 2 6	2 4 6	4 2 6

Solutions are above (and their reversals).

Abstracting, the method we used above is:

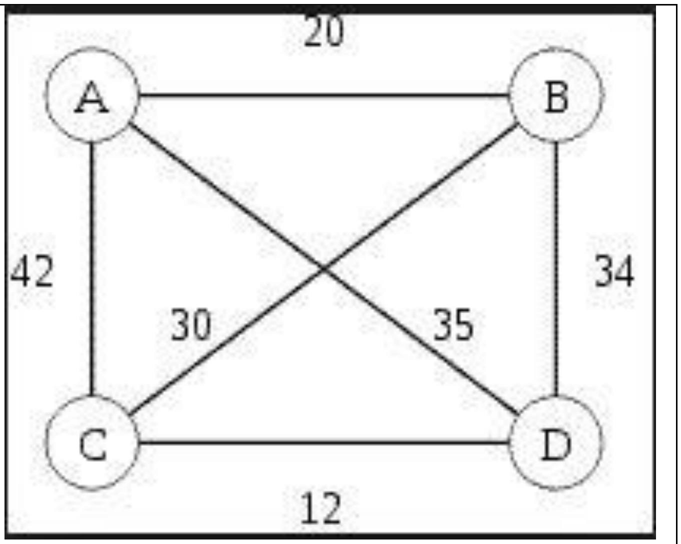
- input fragment sizes $\{a_1, a_2, \ldots, a_n\}$ ($n-1$ cuts), $\{b_1, b_2, \ldots, b_m\}$ ($m-1$ cuts), and $\{c_1, c_2, \ldots, c_{n+m-2}\}$

- **produce the two position maps:**

$$\mathbf{0}, a_1, a_1 + a_2, \dots, a_1 + a_2 + \cdots + a_n = L$$
$$\mathbf{0}, b_1, b_1 + b_2, \dots, b_1 + b_2 + \cdots + b_m = L$$

- **merge these two and put into order**
- **take the differences above**
- **check whether these differences match $\{c_1, c_2, \dots, c_{n+m-2}\}$**

Above is an example of an Algorithm (p.105), akin to the travelling salesman problem (TSP).



But our method above is not efficient since there are $n! \times m!$ different combinations in step 2 to test. With 1000 EcoRI fragments and 1000 *Hin*dIII (A/AGCTT) fragments, this is astronomically large. An alternate idea is the experimental approach of optical mapping (p.106) or incomplete digestion (p.107); see text for details.

4.5. Building Contigs from Cloned Genome Fragments

4.5.1. How Many Clones Are Needed?: microbial genomes are on the order of $G > 0.5 \times 10^6 bp$ and those for mammals are $G > 10^9 bp$. In contrast, cloning vectors are around $10^4 bp$ for lambda/cosmid vectors and $10^5/10^6 bp$ for BAC/YAC vectors. So, the question arises: how many clones N should there be in a given genomic library?

- **G = genome length (in bp)**
- **L = length of the clone insert (in bp)**
- **f = probability that any chosen base pair is represented in the library**
- **f_c = probability that any chosen base pair is covered after N clones have been drawn**

Next, take one clone: the probability that any particular base is not in the clone (or 'covered by the clone') is $1 - L/G$. After N independent clones, the probability that any chosen base pair is not covered is $(1 - L/G)^N$, so we have:

$$1 - f_c = P(not\ covered, N\ clones) = (1 - L/G)^N$$

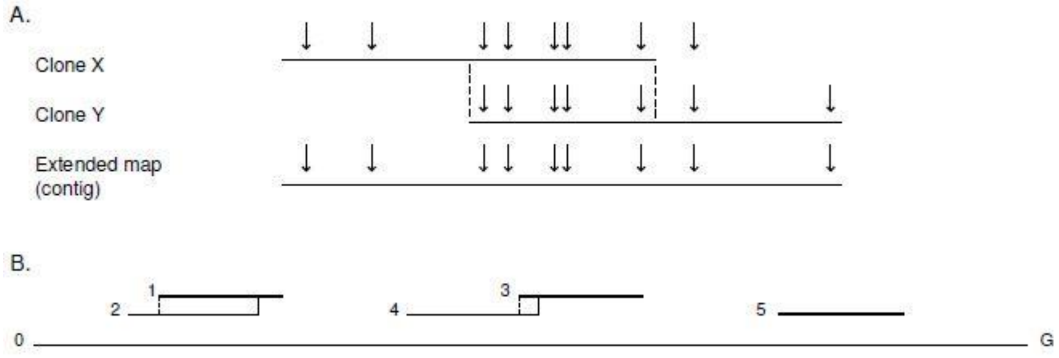
After taking log's, it follows that $N = \log(1 - f_c)/\log(1 - L/G)$

<u>Example</u>: How many cosmid clones are required in a library representing the *E.coli* genome such that $f_c = 0.95$? Here, $G = 4.6 \times 10^6 bp$, $L = 4 \times 10^4 bp$, so $N = \log(0.05)/\log(1 - 4/460) = 343\ clones$. [That's $343\ clones \times 4 \times 10^4 bp\ per\ clone = 13.72 \times 10^6 bp = 2.98G$. So, the coverage, $c = NL/G$, is about 3 here; c is also called the number of genome equivalents (in the clones).]

Further, $1 - f_c = \left(1 - \frac{L}{G}\right)^N = \left(1 - \frac{c}{N}\right)^N \approx e^{-c}$, and $f_c = 1 - e^{-c}$

c	1	2	3	4	5
f_c	0.632	0.865	0.950	0.982	0.993

4.5.2. Building Restriction Maps from Mapped Clones: in Chap. 8, we study the "bottom-up" approach, which is developing a larger segment of the map from data for smaller segments. If the inserts in two different clones contain some of the same restriction sites, then the inserts may share a region in common (they overlap), and a larger, contiguous mapped region can be recognized by extending the physical map on both sides of the region of overlap. A contig is a genome segment represented by two or more overlapping clones (see Fig. 4.3A).



How rapidly will the map approach completion as coverage c increases, and when should we stop picking clones at random and move to directed gap closure?

4.5.3. Progress in Contig Assembly: as seen in Fig.4B, some overlaps are too short to be recognized (3 and 4), whereas other clones overlap enough to be recognized (1 and 2); here, clones 3

and 4 form a contig that cannot be recognized and clone 5 is a singleton. Contigs (recognized and unrecognized) and singletons constitute "islands" in the "sea" of genomic DNA (the genome).

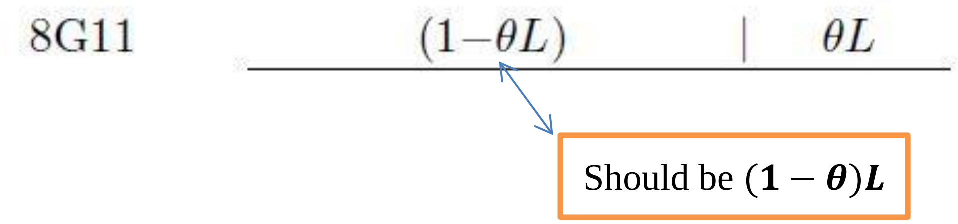
- $\boldsymbol{N}$ **= number of clones**
- $\boldsymbol{L}$ **= length of the insert in each clone**
- $\boldsymbol{G}$ **= genome length**
- $\boldsymbol{\Omega}$ **= minimum amount of overlap required for detection of that overlap**
- $\boldsymbol{\theta}$ **= fraction of $\boldsymbol{L}$ corresponding to $\boldsymbol{\Omega}$ … so that $\boldsymbol{\Omega = \theta L}$**

We start by looking at the distribution of clone inserts along the genome (islands must start with one particular clone), and specifically looking at the left endpoint of this clone (the remaining $\boldsymbol{L-1}$ bp are then to the right of this point). Also, we know that $\boldsymbol{N}$ clones have been drawn, so we know that the probability that any genome position corresponds to the left end of a clone is $\boldsymbol{\lambda = N/G = c/L}$. [This follows since we let $\boldsymbol{Y_N}$ be the RV corresponding to the # of left endpoints, and $\boldsymbol{\lambda}$ be the probability that any one bp is a left endpoint. Then, $\boldsymbol{E(Y_N) = G\lambda = N}$.] Hence, in an interval of length $\boldsymbol{x}$, if $\boldsymbol{K}$ is the number of left endpoints in this interval, then

$$\boldsymbol{P(K = k\ in\ this\ interval) = \frac{(xc/L)^k}{k!} e^{-xc/L}}$$

What about the number of apparent islands? We use the term "apparent" since clone pairs that overlap but not enough to be detected will be counted as two islands instead of one. We'll enumerate the number of islands of clones by counting their right endpoints, which is the same as the number of clones that are at

the right end of an island. Γ is the expected number of islands, so that $\Gamma = N \times P(a\ clone\ is\ at\ the\ right\ end\ of\ an\ island).$ To understand, consider the following clone insert (named 8G11)



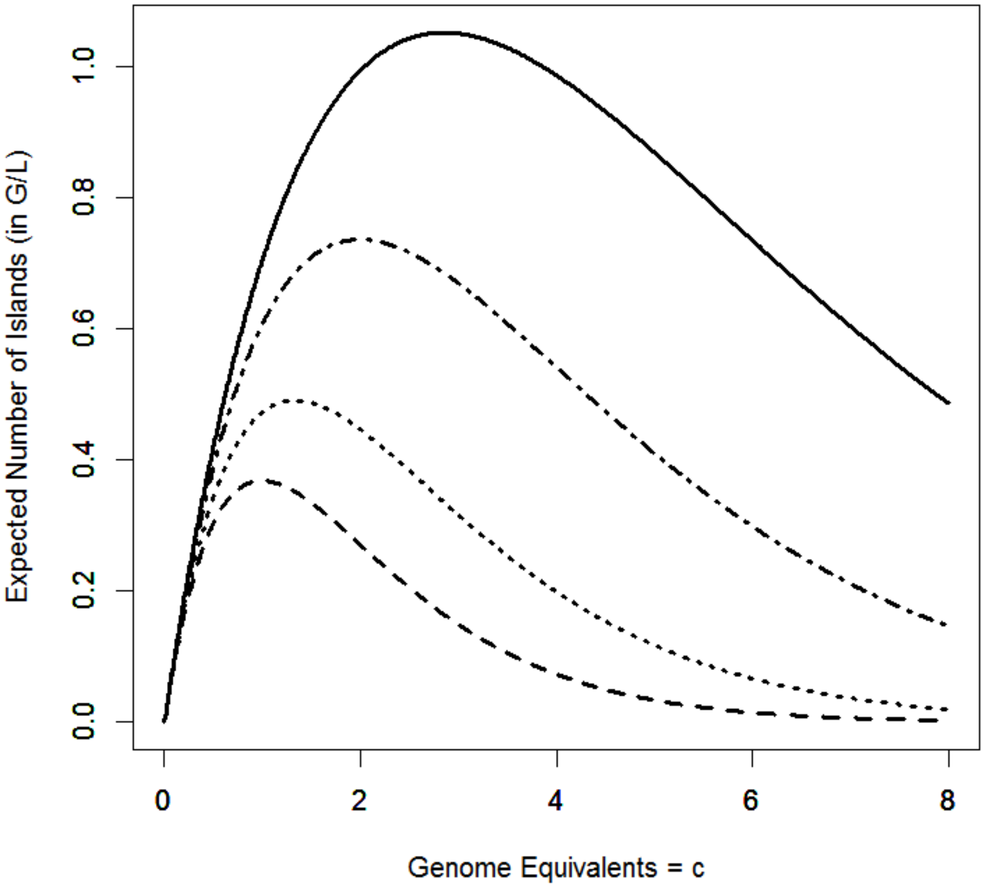
What is the probability that clone 8G11 is at the right end of an island? If there is another clone whose left end lies within θL of the right end of clone 8G11, we can’t tell. But if there is another clone whose left end lies within the segment of length $(1-\theta)L$, we are then certain that 8G11 is not the right endpoint of the island; so we want no left endpoints in the $(1-\theta)L$ portion; this corresponds to $k = 0$ and $x = (1-\theta)L$ above. Hence, the expected number of islands is therefore:

$$\Gamma = Ne^{-c(1-\theta)} = \frac{G}{L}\left(ce^{-(1-\theta)c}\right) = \frac{G}{L}f(c)$$

The expression is sufficient to calculate; the rest is to express in units of $\frac{G}{L}$. Also, $f(c)$ is plotted below versus c and for various θ.

```
eqn.4.11<-function(c,th) c*exp(-(1-th)*c)
c<-seq(0,8,length=1000)
y00<-eqn.4.11(c,0)
y25<-eqn.4.11(c,0.25)
y50<-eqn.4.11(c,0.50)
y65<-eqn.4.11(c,0.65)
yrange<-range(y00,y25,y50,y65)
```

```
xrange<-range(c)
plot(c,y00,type="l",lwd=3,lty=2,xlim=xrange,ylim=yrange,xlab="",ylab="")
par(new=T)
plot(c,y25,type="l",lwd=3,lty=3,xlim=xrange,ylim=yrange,xlab="",ylab="")
par(new=T)
plot(c,y50,type="l",lwd=3,lty=4,xlim=xrange,ylim=yrange,xlab="",ylab="")
par(new=T)
plot(c,y65,type="l",lwd=3,lty=1,xlim=xrange,ylim=yrange,
xlab="Genome Equivalents = c",ylab="Expected Number of Islands (in G/L)")
```



In the above, remember what's given by the researcher are G, θ and usually L: we can choose the c (i.e., the N since $c = NL/G$).

Why do the curves increase and then decrease? "When the mapping is started, there are no islands. At the beginning, as we draw each new clone for mapping, it is more likely to be a singleton (at that stage) than to overlap a clone already drawn, so the number of islands begins to climb. As the number of clones examined increases, existing contigs begin to increase in length as clones overlap their ends. So, later in the process, an ever-increasing proportion of the genome is being spanned by contigs. Any new clones that fall inside an existing contig do not add to the number of contigs or islands. Some clones will overlaps the left end of one contig and the right end of another, causing the contigs to merge. This means that the number of contigs and islands begins to drop later in the process…"

Finally, let's estimate the number of singletons: a given clone (e.g. 8G11) is a singleton if no other clone's L endpoint falls in the left $(1-\theta)L$ portion of 8G11 AND no other clone's R endpoint falls in the right $(1-\theta)L$ portion of 8G11. These are independent events and the lengths are the same, so we get

$$E(\#\ of\ singletons) = Ne^{-2(1-\theta)c}$$

Computational Example 4.2: *E-coli* and lambda cloning vectors give: $N = 1025, L = 1.55 \times 10^4, G = 4.7 \times 10^6$ We wish to find the expected number of islands and singletons and compare with the actual values. Here, finding θ takes some work (see below), and $c = 15.8875 \times 10^6 / 4.7 \times 10^6 = 3.3803$. Let x be the expected number of cuts (cleaved sites) in a clone, then:

$$\frac{9700}{4.7 \times 10^6} = \frac{x}{1.55 \times 10^4}, so\ x \approx 32$$

It is also stated that 6 consecutive sites are required to declare an overlap, so $\boldsymbol{\theta \approx \frac{6}{32} \approx 0.19}$. It follows that the expected number of islands is $\boldsymbol{\Gamma = (303.226)(3.3803)e^{-2.84} = 66.3}$, and the expected number of singletons is $\boldsymbol{1025 \times e^{-2(1-0.19)3.38} = 4.3}$. Since the actual values were 70 and 7 respectively, the match here is excellent.

4.6. Minimal Tiling Clone Sets and Fingerprinting:

Here, we don't need to generate a complete restriction map, just a minimum number to help reconstruct the genome sequence – this is called the minimal tiling clone set (see Figure 4.3C).

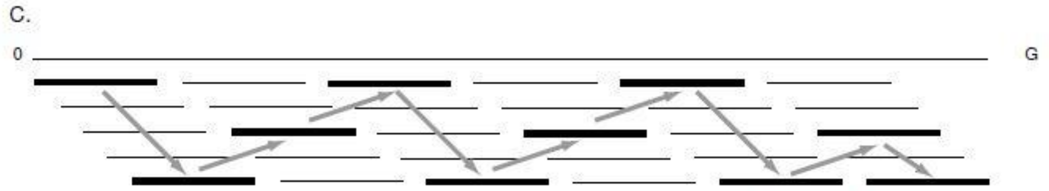


Fig. 4.3. Mapping a large genome by assembling maps of representative clones. Panel A: Clones X and Y are seen to overlap because they share some of the same restriction fragments. This allows their maps to be joined to form a **contig**. Panel B: Cloned inserts can overlap sufficiently for that overlap to be detected (1 and 2), may overlap insufficiently for overlap detection (3 and 4), or may be singletons (5) having no context information supplied by neighbors. Contigs, singletons, and selected clones with undetected overlaps are called **islands** (see the text). Panel C: Once multiple clone overlaps have been constructed such that they span a whole genome or genome segment, there will be several clones covering each region. In this illustration, coverage $c \sim 5$. To reduce the redundancy, a subset of minimally overlapping clones is chosen (heavy lines) that still spans the whole genome. The path through the graph representing these clones (arrows) is called a *minimal tiling path*.

A simplified example is given on pp.116-7 reducing the number of needed clones from 3 to 2, but in general the process isn't difficult – "… it is relatively easy to determine the sequences for about 500 bp into the insert from each end of the segment clones. With this sequence, we can design PCR primers that will amplify DNA between them wherever that sequence is present and, in particular, within the inserts in overlapping clones. We would then start with clone X and design primers for its right end. By using appropriate pooling schemes, it is relatively easy to identify

other clones that overlap the right end of X because an amplified product can be produced from the appropriate clone pools. Those are the only clones that need to be fingerprinted to determine the ones that have minimal overlap with the right end of X. If the left end of Y overlaps the right end of X, then we can design primers for the right end of Y, screen the library for clones that overlap its right end, fingerprint those clones, and continue the process until we have produced a minimal tiling path clone set. What we have just described is equivalent to employing sequence-tagged connectors to generate a minimal tiling clone set like those used for clone-by-clone shotgun sequencing (Chapter 8).”