Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

Human Population Genomics

Outline

1. Statistical Distributions in Genome Sequence Assembly
    - Island Statistics
    - Small Islands
2. Finger Prints and Maps

Summary of the lecture (Mon February 9, 2009)

1. Lecture notes are available on the course website.
2. Please feel free to download the slides from the website.

“I believe that ... *propinquity of descent*, — the only known cause of the similarity of organic beings, — is the bond, hidden as it is by various degrees of modification, which is partially revealed to us by our classifications.”

Charles Darwin; *On the Origin of Species by Means of Natural Selection*; 1859.

Outline

1. Statistical Distributions in Genome Sequence Assembly
   - Island Statistics
   - Small Islands
2. Finger Prints and Maps

Notations

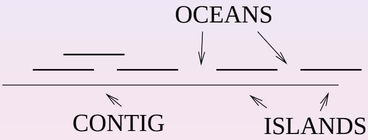
Consider a genome of length G that has been uniformly randomly sampled to collect N clones each one of length L. The parameters of interest are now extended to include the overlap threshold:

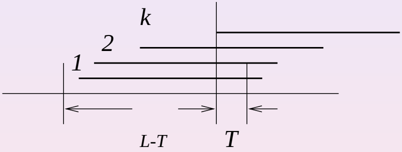
$$\begin{aligned}
G &= \text{Genome length (in bp).} \\
L &= \text{Length of a clone.} \\
N &= \text{Number of clones.} \\
\alpha = \left(\frac{N}{G}\right) &= \text{Expected \# clones starting in a unit interval of } G \\
&= \text{Probability of a clone starting at a given site} \\
c = \left(\frac{LN}{G}\right) &= \mathit{Coverage} = L\alpha
\end{aligned}$$

Notations (Contd)

$$
\begin{aligned}
T &= \textit{Overlap parameter} \\
&= \text{\# base pairs two clones must have in common to ensure their overlap.} \\
\theta = \left(\frac{T}{L}\right) &= \textit{Overlap threshold ratio} \\
\sigma &= 1 - \theta
\end{aligned}
$$

$$L - T = L(1-\theta) = L\sigma = \frac{c\sigma}{\alpha}.$$


OCEANS
CONTIG
ISLANDS


k
2
1
L-T
T

There are number of simple conclusions that one can make:

- The probability that an island begins at *a*

$$I_a = \mathbb{P}r[V_a = 0 \ \ \& \ \ \exists\, i \ \ S_{i,a} = 1] = \alpha e^{-c\sigma}.$$

- The expected number of islands (= expected number of oceans =)

$$\sum_{a=1}^{G} I_a = G\alpha e^{-c\sigma} = Ne^{-c\sigma}.$$

- Thus if we choose $c = \ln G/(1-\theta)$ and thus make the effective total length of the clones

  $$N(L-T) = NL(1-\theta) = Gc\sigma = G \ln G$$

  then the expected number of contigs is 1 with high probability, assuming that $\ln G < L\sigma$.
- Another way of saying the same would be that we must make

  $$\theta \leq \max(1-(\ln G/c), 0),$$

  if we wish to get a genome wide complete map.
- For instance, if we have a 46 × coverage clone library for human (as claimed by Celera), then we need to use a $\theta \leq 0.474$.

- The probability that the *i*th clone begins (or, symmetrically, ends) an island is

$$\mathbb{P}r[\exists\, a \;\; S_{i,a} = 1 \;\, \& \;\, V_a = 0] = e^{-c\sigma}.$$

- The probability that an island has exactly $j+1$ clones

$$Z_{I,j+1} = \left(1 - e^{-c\sigma}\right)^j e^{-c\sigma} \approx e^{-(c\sigma + je^{-c\sigma})}.$$

- Thus the probability that an island is a singleton is $e^{-c\sigma}$ and the probability that it is a non-trivial contig is $1 - e^{-c\sigma}$.

- The expected number of singleton islands

  $$Ne^{-2c\sigma},$$

  and the expected number of contigs is

  $$Ne^{-c\sigma} - Ne^{-2c\sigma}.$$

- The expected number of clones per island is then simply

  $$\bar{j} = e^{c\sigma}.$$

- Suppose an apparent island ends at position y. What is the probability that there is an ocean of length exactly x starting at y? This is simply

$$\begin{aligned} &\mathbb{P}r[\text{ No clone starts in the interval } [y-T, y+x] \\ &\qquad \text{and a clone starts at } x+1] \\ &= \alpha(1-\alpha)^{x+T} \\ &\approx e^{-\alpha T}\alpha e^{-\alpha x} \\ &= e^{-c\theta}\alpha e^{-\alpha x}. \end{aligned}$$

- Since the moment generating function in this case is

$$\Psi(t) = \frac{\alpha e^{-c\theta}}{\alpha - t},$$

the expected length of an ocean in base pairs is

$$\mathbb{E}[X] = \Psi'(0) = \frac{e^{-c\theta}}{\alpha} = \frac{L}{c} e^{-c\theta},$$

and the variance is

$$\begin{aligned} \mathbb{V}ar[X] &= \Psi''(0) - (\Psi'(0))^2 = \frac{e^{-c\theta}(2 - e^{-c\theta})}{\alpha^2} \\ &= \frac{L^2 e^{-c\theta}(2 - e^{-c\theta})}{c^2}, \end{aligned}$$

and

$$\text{Std. Dev.}[X] = \frac{L}{c} e^{-c\theta/2} \sqrt{(2 - e^{-c\theta})}$$

- Note also that the expected fraction of the genome in the oceans (i.e., not represented by the clones) is Ge^{-c} and the total number of oceans is $Ne^{-c\sigma}$. Thus the expected length of an ocean is

$$\frac{Ge^{-c}}{Ne^{-c\sigma}} = \frac{Ge^{-c(1-\sigma)}}{N} = \frac{Ge^{-c\theta}}{N} = \frac{L}{c}e^{-c\theta}.$$

- Thus the probability that an ocean is of length greater than $N(2\ln N - c)/G$ is

$$\begin{aligned} e^{-c\theta} \int_{\alpha(2\ln N - c)}^{\infty} & e^{-\alpha x}\ \alpha\ dx \\ &= e^{-c\theta} e^{-(2\ln N - c)} \\ &= \frac{e^{c\sigma}}{N^2}. \end{aligned}$$

- Since the expected number of oceans is $Ne^{-c\sigma}$, the probability that all the oceans are of length smaller than $N(2\ln N - c)/G$ is

$$\left(1 - \frac{e^{c\sigma}}{N^2}\right)^{Ne^{-c\sigma}} \approx e^{-(1/N)},$$

  very close to 1, for large N.
- In particular, if

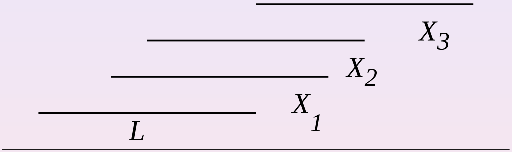
$$\frac{2\ln N}{N} \leq \frac{L}{G},$$

  then $2\ln N - c \leq 0$ and all oceans are of length 0 almost surely, and the contigs cover almost all of the genome.

- Let us try to estimate the expected length of an island in base pairs, with the following heuristic arguments. The expected length of all the oceans is $[(L/c)e^{-c\theta}][Ne^{-c\sigma}] = Ge^{-c}$.
- Thus the "total length" of all the islands (of course, without properly accounting for the undetected overlaps among the islands) is
$$G - Ge^{-c} = G(1 - e^{-c}),$$
and the expected "length" of an island in base pairs is
$$\begin{aligned} \frac{G(1-e^{-c})}{Ne^{-c\sigma}} &= \frac{G}{N}\left(\frac{1-e^{-c}}{e^{-c\sigma}}\right) = \frac{L}{c}(e^{c\sigma} - e^{-c\theta}) \\ &\approx L\left(\frac{e^{c\sigma} - 1 + c\theta}{c}\right) = L\left(\frac{e^{c\sigma}-1}{c} + \theta\right). \end{aligned}$$

- For small θ, the above expression is correct, but may need to be modified appropriately, if we wish to account for significantly larger θ and hence the unaccounted for overlaps among the apparent islands.
- Interestingly enough, for $\theta = 1$ (thus, $\sigma = 0$), the above expressions yields for the expected length of an island a value of *L*, which is in fact the correct value!
- We will see soon (using a more detailed analysis) that the above expression is correct for all values of θ.

1. Let $0 = s_1, s_2, \ldots, s_a, s_{a+1}, \ldots, s_{a+k}, s_{a+k+1}, \ldots$, etc., be the left ends of the clones given in the increasing order.
2. Let s_a, be the left end of the first clone of an island of interest. Let X_i be a random variable denoting $s_{a+i} - s_{a+i-1}$, the distance from the left end of the *i*th clone to the left end of the next clone in the same island, if there is one such clone.
3. Of course $X_i = L$, if in fact it is the last clone in the island and the next clone starts a new island.


X_3
X_2
X_1
L

More formally, let $x_i = s_{a+i} - s_{a+i-1}$

$$X_i = \begin{cases} x_i, & \text{if } 0 < x_i < L\sigma; \\ L, & \text{otherwise.} \end{cases}$$

Thus the length of the island is a random variable given as

$$\Lambda = X_1 + X_2 + \cdots + X_k,$$

such that $x_1 < L\sigma$, ..., $x_{k-1} < L\sigma$ and $x_k \geq L\sigma$. Thus $\mathbb{E}[\Lambda]$ is the expected length of an apparent island.

Martingale Theory

Definition

A stochastic process $\{Z_n, n \geq 1\}$ is said to be a *martingale* process if

$$\mathbb{E}[|Z_n|] < \infty, \quad \text{for all } n$$

and

$$\mathbb{E}[Z_{n+1}|Z_1, Z_2, \ldots, Z_n] = Z_n.$$

1. Taking expectations of the last equation, we get

$$\mathbb{E}[Z_{n+1}] = \mathbb{E}[Z_n], \quad \Rightarrow \quad \forall\, n\ \mathbb{E}[Z_n] = \mathbb{E}[Z_1].$$

2. A classical example of a martingale is based on the fact that the partial sums of independent random variables having mean 0 is a martingale.

1. Let X_i's be any random variables, for instance, the ones defined for the clones in an island.
2. The random variables

   $$X_i - \mathbb{E}[X_i | X_1, \ldots, X_{i-1}]$$

   have mean zero and their partial sums

   $$Z_n = \sum_{i=1}^{n} \{X_i - \mathbb{E}[X_i | X_1, \ldots, X_{i-1}]\}$$

   form a martingale assuming that the expectations of its magnitudes are finite.
3. It is a martingale of mean zero.

- Note that

$$Z_{n+1} = Z_n + X_{n+1} - \mathbb{E}[X_{n+1}|X_1, \ldots, X_n]$$

  and taking the expectations on both sides, we have

$$\begin{aligned}\mathbb{E}[Z_{n+1}|Z_1, \ldots, Z_n] &= Z_n + \mathbb{E}[X_{n+1}|X_1, \ldots, X_n] - \mathbb{E}[X_{n+1}|X_1, \ldots, \\ &= Z_n.\end{aligned}$$

- Furthermore

$$\mathbb{E}[Z_{n+1}] = \mathbb{E}[Z_1] = \mathbb{E}[X_1] - \mathbb{E}[X_1] = 0.$$

- Note, further, that if X_i's are independent (as in our case) then the definition above simplifies to

$$Z_n = \sum_{i=1}^{n} \{X_i - \mathbb{E}[X_i]\}$$

Martingale Stopping Theorem

Definition

For the process $\{X_n, n \geq 1\}$, the integer valued positive random variable N is said to be a *random time*, if the event $\{N = n\}$ is determined by the random variables X_1, X_2, ..., X_n. If

$$\mathbb{P}r[N < \infty] = 1,$$

then the random time N is said to be a *stopping time*.

- It follows from the *Martingale Stopping Theorem*, that if N is a stopping time for a martingale $\{Z_n, n \geq 1\}$ such that (1) $\mathbb{E}[N] < \infty$ and

$$(2) \exists\, M < \infty \quad \mathbb{E}[|Z_{n+1} - Z_n| : Z_1, \ldots, Z_n] < M$$

then $\mathbb{E}[Z_N] = \mathbb{E}[Z_1]$.

Thus we have the following corollary:

Corollary

Let X_i $(i \geq 1)$ be independent and identically distributed random variables with a stopping time N, $\mathbb{E}[N] < \infty$ and $\mathbb{E}[|X|] < \infty$, then

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\mathbb{E}[X]. \quad \square$$

Proof.

First define $Z_n = \sum_{i=1}^n X_i - \mathbb{E}[X_i]$ as before. Then

$$\begin{aligned}\mathbb{E}[|Z_{n+1} - Z_n| : Z_1, \ldots, Z_n] &= \mathbb{E}[|X_{n+1} - \mathbb{E}[X]|] \\ &\leq \mathbb{E}[|X|] + |\mathbb{E}[X]| < 2\mathbb{E}[|X|] \\ &< \infty.\end{aligned}$$

Thus $\mathbb{E}[Z_N] = \mathbb{E}[Z_1] = 0$, and

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^N X_i - \mathbb{E}[X_i]\right] &= \mathbb{E}\left[\sum_{i=1}^N X_i\right] - \mathbb{E}\left[\sum_{i=1}^N \mathbb{E}[X]\right] \\ &= \mathbb{E}\left[\sum_{i=1}^N X_i\right] - \mathbb{E}[N]\mathbb{E}[X] = 0,\end{aligned}$$

as needed. □

Expected Length of an Island

- Now, we are ready to estimate the expected length of an island.
- Recall that, by definition X_i's are i.i.d. and are defined as

$$X = \begin{cases} x, & \text{if } 0 < x < L\sigma; \\ L, & \text{otherwise.} \end{cases}$$

where x denotes the position where the next clone starts, and is distributed as $(1-\alpha)^x \alpha \approx \alpha e^{-\alpha x}$.

Thus,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{L\sigma} x\alpha e^{-\alpha x}\, dx + \int_{L\sigma}^{\infty} L\alpha e^{-\alpha x}\, dx \\
&= \frac{1}{\alpha}\int_0^{c\sigma} ye^{-y}\, dy + L\int_{c\sigma}^{\infty} e^{-y}\, dy \\
&= \frac{L}{c}\left[-ye^{-y}\Big|_0^{c\sigma} - e^{-y}\Big|_0^{c\sigma}\right] + L\left[-e^{-y}\Big|_{c\sigma}^{\infty}\right] \\
&= -L\sigma e^{-c\sigma} + \frac{L}{c}\left(1 - e^{-c\sigma}\right) + Le^{-c\sigma} \\
&= L\left(\frac{1 - e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right)
\end{aligned}$$

Also, note that $0 < X \leq L$ and $\mathbb{E}[|X|] < L$.

1. Let k be the stopping time denoting that the kth clone in the island terminates the island. It is easily seen that it is the stopping time as $X_1 < L\sigma, \ldots, X_{k-1} < L\sigma$ and $X_k = L$.
2. Now the probability that a given $x_j < L\sigma$, and thus $j < k$, is given by

$$\begin{aligned}
\mathbb{P}r[x < L\sigma] &= \int_0^{L\sigma} \alpha e^{-\alpha x}\, dx \\
&= \int_0^{c\sigma} e^{-y}\, dy = \left[-e^{-y} \Big|_0^{c\sigma} \right] = 1 - e^{-c\sigma} < 1.
\end{aligned}$$

Thus the $\mathbb{E}[k] = \sum_{j=1}^{\infty} j(1 - e^{-c\sigma})^{j-1} e^{-c\sigma} = e^{c\sigma} < \infty \ldots$

1. And the expected length of an island is

$$\begin{aligned}\mathbb{E}[k]\mathbb{E}[X] &= e^{c\sigma}\, L\left(\frac{1-e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right)\\ &= L\left(\frac{e^{c\sigma}-1}{c} + \theta\right).\end{aligned}$$

2. Rather familiar!
3. For large c, we can use the following approximation:

   Expected length of an island is $\approx Le^{c\sigma}/c$—recall that the expected length of an ocean is $Le^{-c\theta}/c$. In particular, their ratio is $\approx e^{c}$, giving an idea about the exponential growth in the sizes of the islands with the coverage.

1. We saw that an island is on the average of length $\approx Le^{c\sigma}/c$.
2. Here, we will see that actually this estimate is quite sharp, in the sense almost all islands are of length larger than $Le^{3c\sigma/4}[1-o(1)]/c$.
3. First note that the probability that an island has at least k' clones is
$$(1-e^{-c\sigma})^{k'} \approx e^{-k'e^{-c\sigma}}.$$
4. Thus
$$\begin{aligned} & \mathbb{P}r[\text{ An island has more than } k' = e^{3c\sigma/4} \text{ clones }] \\ \geq \quad & e^{-e^{3c\sigma/4}e^{-c\sigma}} \\ = \quad & e^{-e^{-c\sigma/4}}. \end{aligned}$$

1. Now, if an island has k' or more clones then, we know that its length is bigger than

$$\begin{aligned}\Lambda' &= X_1 + X_2 + \cdots + X_{k'} \\ &= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_{k'}] + Z_{k'} \\ &= Le^{3c\sigma/4}\left(\frac{1-e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right) + Z_{k'}.\end{aligned}$$

2. Recall that $\{Z_n, n \geq 1\}$ is a martingale with mean zero. $Z_0 = 0$. $Z_n - Z_{n-1} = X_n - \mathbb{E}[X]$ and since $0 \leq X_n \leq L\sigma$, we have, for all n, $-L \leq -\alpha \leq Z_n - Z_{n-1} \leq \beta \leq L$ with $\alpha + \beta = L\sigma$.
3. By the *generalized Azuma inequality*, we have for any positive c (e.g., $c = L\sigma e^{-c\sigma/4}$)

$$\mathbb{P}r[Z_{k'} \leq -k'c] \leq e^{-2k'c^2/(\alpha+\beta)^2}.$$

- Thus the probability that $Z_{k'} \leq -e^{3c\sigma/4}(L\sigma e^{-c\sigma/4})$, is less than
$$e^{-2(e^{3c\sigma/4})(L^2\sigma^2 e^{-c\sigma/2})/(L^2\sigma^2)} = e^{-2e^{c\sigma/4}}.$$
- Thus with probability close to 1, we have
$$Z_{k'} \geq -\frac{Le^{3c\sigma/4}}{c}(c\sigma e^{-c\sigma/4}).$$

- Thus with probability close to 1, we have

$$\begin{aligned}\Lambda' &\geq \frac{Le^{3c\sigma/4}}{c}\left(1 - e^{-c\sigma} + c\theta e^{-c\sigma} - c\sigma e^{-c\sigma/4}.\right)\\ &= \frac{Le^{3c\sigma/4}}{c}(1 - o(1)).\end{aligned}$$

- And, this also implies that almost all the islands are of length larger than this.

Theorem

Consider N clones each of length L from a genome of length G constructed by sampling these uniformly over the genome. Let c denote their coverage $c = NL/G$ and θ denote an overlap threshold denoting the fractions of length that two clones must overlap in order for the overlap to be detected. The followings summarize our earlier results.

- *The expected number of apparent islands and oceans is $Ne^{-c(1-\theta)}$. The expected number of singleton islands is $Ne^{-2c(1-\theta)}$.*
- *The expected number of clones in an island is $e^{c(1-\theta)}$.*

Theorem (Contd.)

- *The expected length of an island is*

$$L\left(\frac{e^{c(1-\theta)}-1}{c}+\theta\right).$$

- *The expected length of an ocean is*

$$\frac{Le^{-c\theta}}{c}.$$

Theorem (Contd.)

- *An island is of length larger than*

$$(1+o(1))\frac{L}{c}e^{3c(1-\theta)/4},$$

almost surely.

- *An ocean is of length smaller than*

$$\max\left[0, \frac{L}{c}(2\ln(G/L) + 2\ln c - c)\right],$$

almost surely. □

Outline

1. Statistical Distributions in Genome Sequence Assembly
   - Island Statistics
   - Small Islands

2. Finger Prints and Maps

Mapping

- We start with the concept of finger prints and maps of a clone or a genome.
- For our purpose here, a clone is a rather loosely defined object...
- More precisely, it denotes a 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

- We may wish to characterize these clones in many ways — without ever sequencing them...
- We may sequence just the ends to create “mated pairs;” we may hybridize probes to them to make “probe maps;” we may cut them with restriction enzymes to make “restriction maps.”

- For instance, if we take a 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*.
- This information can be used (*probabilistically*) to detect, if two clones overlap!

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**.

- 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.

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

and

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

- 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.

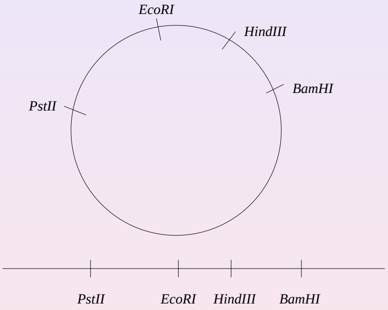
Hae III:	**GGCC|GGCC** **CCGG|CCGG**	Blunt end 8-cutter
*Eco*R I:	**G|AATTC** **CTTAA|G**	Sticky end 6-cutter

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).

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_u**GCCG**$\|P_y$ $P_y\|$**CGGC**P_u
Haemophilus aegyptius	*Hae* III	**GG|CC** **CC|GG**

MicroOrganism	Restriction Enzyme	Restriction Site
Haemophilus influenza Rd	*Hind* II	$\mathbf{GT}P_y\|P_u\mathbf{AC}$ $\mathbf{CA}P_u\|P_y\mathbf{TG}$
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**


EcoRI
HindIII
BamHI
PstII
PstII
EcoRI
HindIII
BamHI

- 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. $p_k = 1/4^k$ and if a clone is of length L, we expect it to have $n = p_k L$ fragments.
- 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. This accuracy will be measured by a parameter β.
- 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. The partial digestion rate will be measured by a cutting efficiency p_c.

- Finally, the spacing between the markers is directly related to how informative the map is and determines its *resolution*.
- 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.

Three Types of Restriction Maps

Finger Print: All the restriction fragments are known, but their order information is lost.

Incomplete Maps: Most of the restriction fragments are known, together with their order information.

Complete Maps: All of the restriction fragments are known, together with their order information.

A particularly important parameter to consider is the effective coverage $\bar{c} = c(1-\theta)$ that could be accomplished by the three different kinds of mapping...

Comparing the effective coverage under three different overlapping techniques (i.e., finger print, incomplete maps and complete maps), we see that

Finger Print:

$$\bar{c}_{fp} = c\left(1 - \frac{3\beta n}{2}\right).$$

Incomplete Maps:

$$\bar{c}_{IM} = c\left(1 - \frac{3\beta}{2p_c^3}\right).$$

Complete Maps:

$$\bar{c}_{CM} = c\left(1 - \frac{K}{n}(1 - \beta/2)\right).$$

[End of Lecture #4]