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
2. Finger Prints and Maps

Summary of the lecture (Mon February 23, 2009)

1. Let’s talk a bit about the projects...
2. Ideas
3. Teams
4. Grading?

"To the editor of *Science*:
I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists."
G.H. Hardy, *Mendelian Proportions in a Mixed Population*, SCIENCE, **28**:49–50, 1908.

Outline

1. Statistical Distributions in Genome Sequence Assembly

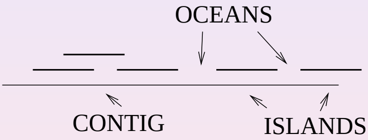
2. Finger Prints and Maps

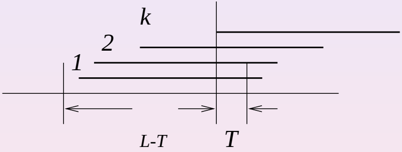
$$
\begin{aligned}
G &= \text{Genome length (in bp).}\\
L &= \text{Length of a clone.}\\
N &= \text{Number of clones.}\\
\alpha = \left(\frac{N}{G}\right) &= \text{Expected \# clones starting in a unit interval of } G\\
&= \text{Probability of a clone starting at a given site}\\
c = \left(\frac{LN}{G}\right)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

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

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

2 Finger Prints and Maps

Mapping

- We start with the concept of finger prints and maps of a clone or a genome.
- A clone is 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

- 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*.

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.

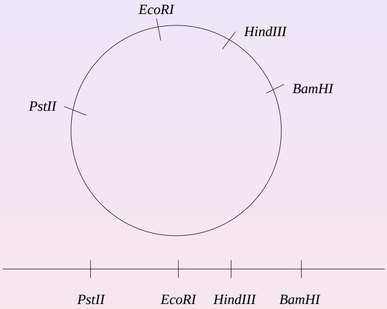
Hae III:	**GGCC|GGCC** **CCGG|CCGG**	Blunt end 8-cutter
*Eco*R I:	**G|AATTC** **CTTAA|G**	Sticky end 6-cutter

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	$\mathbf{G}\vert\mathbf{GATCC}$ $\mathbf{CCTAG}\vert\mathbf{G}$
Brevibacterium albidum	*Bal* I	$\mathbf{TGG}\vert\mathbf{CCA}$ $\mathbf{ACC}\vert\mathbf{GGT}$
Escherichia coli RY13	*Eco* RI	$\mathbf{G}\vert\mathbf{AATTC}$ $\mathbf{CTTAA}\vert\mathbf{G}$
Haemophilus aegyptius	*Hae* II	$P_u\mathbf{GCCG}\vert P_y$ $P_y\vert\mathbf{CGGC}P_u$
Haemophilus aegyptius	*Hae* III	$\mathbf{GG}\vert\mathbf{CC}$ $\mathbf{CC}\vert\mathbf{GG}$

MicroOrganism	Restriction Enzyme	Restriction Site
Haemophilus influenza Rd	*Hind* II	$\mathbf{GT}P_y \mid P_u\mathbf{AC}$ $\mathbf{CA}P_u \mid P_y\mathbf{TG}$
Haemophilus influenza Rd	*Hind* III	$\mathbf{A} \mid \mathbf{AGCTT}$ $\mathbf{TTCGA} \mid \mathbf{A}$
Haemophilus parainfluenza	*Hpa* I	$\mathbf{GTT} \mid \mathbf{AAC}$ $\mathbf{CAA} \mid \mathbf{TTG}$
Haemophilus parainfluenza	*Hpa* II	$\mathbf{C} \mid \mathbf{CGG}$ $\mathbf{GGC} \mid \mathbf{C}$
Providencia stuartii 164	*Pst* I	$\mathbf{CTGCA} \mid \mathbf{G}$ $\mathbf{G} \mid \mathbf{ACGTC}$
Streptomyces albus G	*Sal* I	$\mathbf{G} \mid \mathbf{TCGAC}$ $\mathbf{CAGCT} \mid \mathbf{G}$


EcoRI
HindIII
BamHI
PstII
PstII
EcoRI
HindIII
BamHI

A map can be presented as in the previous 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.

- 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.

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...

NOISE

The process of creating restriction maps (or clone library with restriction enzyme) can be affected by several error processes associated with cleavage.

Partial Digestion: The cleavage processes at all the restriction sites do not proceed to completion.

False Cuts & Star Activity: 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.

Missing Fragments: Some of the small fragments may 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.

Partial Digestion

- The error process (due to partial digestion) 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

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

1. The moment generating function is given by

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

2. Thus

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

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.

False Cuts & Star Activity

- *Star activity* on the other hand corresponds to the loss of restriction enzyme specificity.
  - This occurs 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.

Statistical Distribution of Restriction Sites

- In a genomic DNA of length G, how many restriction fragments does one expect?
- 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 $s \in (\mathbf{A} + \mathbf{T} + \mathbf{C} + \mathbf{G})^+$:

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

Numerically:

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

1. Let $\lambda_k = Gp_k$ (G is the length of a genomic DNA to be cut by a k-cutter restriction enzyme).
2. Thus the number of restriction sites in the DNA has a *Binomial distribution*.

$$\begin{aligned} & \mathbb{P}r[\exists\ M \text{ restriction sites in a DNA of size } G] \\ = \quad & \binom{G}{M} p_k^M (1 - p_k)^{G-M}. \end{aligned}$$

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

1. 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.

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

2. We assume that $p_{i,k} \approx p_k \ll 1$ and are "very weakly" dependent: $p_{i_1,k} p_{i_2,k} \cdots p_{i_l,k} \approx p_k^l$.
3. 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.
4. Now consider the set of all i trials—there are $\binom{G}{i}$ of these, and each i successful trials occur with probability $\approx p_k^i$.

1. Thus

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

Using Brun's sieve, we have:

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

1. 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

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

1. Thus

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

Size of Restriction Fragments

- The number of base pairs in a restriction fragments can be determined as 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

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

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:

Definition (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.” □

- 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?

- Expected length of a restriction fragment $= \mu_k$ — both x, $y \sim \mathrm{Exp}(1/\mu_k)$. $f_X(x) = (1/\mu_k)e^{-x/\mu_k}$.
- The probability that two randomly chosen restriction fragments match (up to $100\beta\%$) :

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

- 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 700*bp* 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.

- Note: Our assumption that all four base pairs have equiprobable and independent occurrences, is often false.
  - 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 0.43 and the dinucleotide **C***p***G** occur with a probability about $1/5$th 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 $G/p_k = 4700/65 \approx 70$.
- 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.
- 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.

Matching Clones

Overlap Detection

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.
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 (while accounting for β: *the relative sizing error*).
- Recall that 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.

Comparing Finger prints

- Let n denote the number of restriction fragments in each clone. Thus the number of fragments in a subregion of length θL 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

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

- Let us agreethat two clones A and B overlap if they have at least k (where $k \leq \lfloor n\theta/2 \rfloor$) restriction fragments in common. If in fact the clones "overlap" then we succeed with probability close to 1.
- How about the converse?

- What is the probability that two randomly chosen unrelated clones may be 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.

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

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

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

- By an application of Brun's sieve, we have

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

  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:

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

  Thus

$$1 \geq \theta \geq \frac{3\beta n}{2} \quad \text{and} \quad \beta \leq \frac{2}{3n}.$$

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” θ must be rather large. This has some serious negative implications!!!

Comparing Restriction Maps

- We agree 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 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

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

- If the false positive probability to be kept below some small error probability ϵ then k should be approximately $(\ln \frac{4}{\epsilon})(1-\beta/2)$.
  Thus, we need the following conditions satisfied:

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

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

Comparing Partially Digested Restriction Maps

- Here, 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. E.g., optical mapping involving genomic DNA molecules from which one can create rough maps with high β and low partial digestion p_c.
- Model the maps by several parameters: relative sizing error, β, 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 θ.

Choose an overlap rule that takes the partial digestion into account. The rule is fairly simple and as follows:

Postulate an alignment, where clone A is aligned with respect to clone B with an overlap ratio bigger than θ (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.

- Note 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

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

- **False positive probability:** 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.
- Thus

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

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

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

- Thus the false positive probability is

- Thus we need to satisfy the following constraints:

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

- Or

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

- 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.

A particularly important parameter to consider is the effective coverage $\bar{c} = c(1 - \theta)$

Then the number of clones in a contig is $e^{\bar{c}}$, which could be considered a good measure of the complexity of the islands.

Comparing the effective coverage under three different overlapping techniques – we have:

1. Finger Print:

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

2. Incomplete Maps:

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

3. Complete Maps:

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

[End of Lecture #5]

THE END