Computational Systems Biology: Biology X

Bud Mishra

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

Human Population Genomics

Outline

Summary of the lecture (Mon February 2, 2009)

1. Questions about the framework based on Crick's Central Dogma? Too rigid? Ignores the dynamics?
   1. Can we analyze the dynamics? Can the key dynamics structures be expressed topologically? Modal logic?
   2. Phylogeny of dynamics? Distance between two dynamics...
   3. Combining genotyping with dynamic traits? eQTL based methodologies?

1. Is population genomics/genetics the correct way to discover genotype-phenotype relationship?
2. What has been tried? What succeeded? What failed? Why? Bad data? Bad algorithms?
3. An in silico laboratory??? What can we learn from it???
   - Reconstructing the evolution;
   - Choosing the most appropriate technology;
   - Choosing the most appropriate algorithms;
   - Combining with other analysis approaches...

In coming to know the Human Genome,
we move nearer to understanding God -

not further away, as science has wrongly
driven us to conclude hitherto; far nearer

to hearing, reading, knowing the Word -
understanding the organic/spirit concept.
–Gillian Ferguson, *In coming to know the Human Genome*

Outline

Brun's Sieve

Theorem

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

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

Proof of Brun's Sieve

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

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

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

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

Proof of Brun's Sieve (Contd)

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

Outline

Whole Genome Sequencing

1. Single Base Accuracy (No substitution, insertion or deletion; No homopolymer problems)
2. Correct Order (No error due to translocation, inversion, rearrangement, etc)
3. No Gaps (Perhaps, except for telomeres and centromere)
4. Hapoltypic
5. Whole Genome
6. Small Amount of Materials; No Amplification (No Colony, PCR, etc.)

Sequencing Technology

1. **Shotgun Approach:** Read long sentences (stretches of DNA); Use overlap information to *assemble* the reads into a genome-wide sequence.
   - Ideally the sentences should be about 0.5 Mb in length;
   - but currently one uses about 500-700 bp reads (with mated-pairs at 10 Kb, 50 Kb and 150 Kb lengths);
   - Cannot tolerate false-positives in overlap (rampant in repeat regions, and unavoidable with haplotypic ambiguities)

Sequencing Technology

1. **Indexing Approach:** Read one base at a time — Each base comes with its location information (hapoltype + location with respect to some unambiguous landmark);
   - Difficult to get long reads (longer than 500 - 700 bps;
   - Cannot tolerate locational errors.

Sequencing Technology

1. **Middle-Way Approach:**
   1. Long words (6 - 8 mers) with imprecise location information or
   2. Short sentences ($\approx$ 100 bps) with long-range "validating" information

Outline

Basics of Lander-Waterman Statistics

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:

$$G = \text{Genome length (in bp).}$$
$$L = \text{Length of a clone.}$$
$$N = \text{Number of clones.}$$
$$c = \frac{LN}{G} = \textit{Coverage}.$$

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

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

Hence, by Brun's sieve, we have:

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

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

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

More Notations

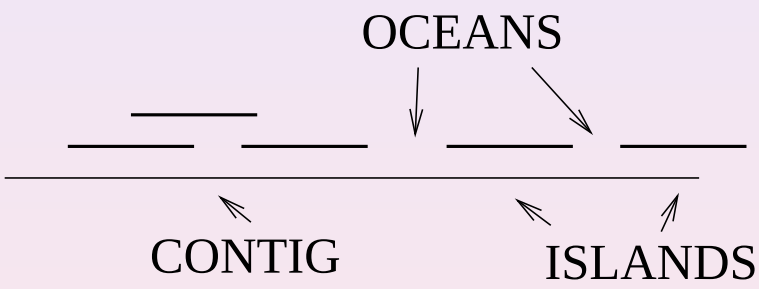
Consider a genome of length G that has been uniformly randomly sampled to collect N clones each one of length L. The parameters of interest are now extended to include the overlap threshold:

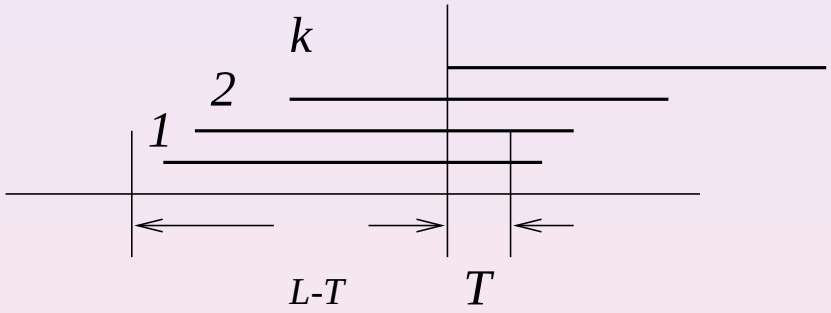
$$
\begin{aligned}
G &= \text{Genome length (in bp).} \\
L &= \text{Length of a clone.} \\
N &= \text{Number of clones.} \\
\alpha = \left(\frac{N}{G}\right) &= \text{Expected \# clones starting in a unit interval of } G \\
&= \text{Probability of a clone starting at a given site} \\
c = \left(\frac{LN}{G}\right)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

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

- In order to understand the finer structure of the overlapping clones, we need to consider few statistical properties of the "*islands*" and "*oceans*."

Definition

An (apparent) *island* is defined to be a maximal set of clones that are closed under the reflexive and transitive closure of the relation induced by the overlap rule.

- Thus an apparent island will always cover a connected subinterval of the genome; however, note that it may still be possible that the union of several islands may also cover a connected subinterval.
- This happens because even when two islands overlap, if they overlap over a portion that is smaller than the overlap parameter T such an overlap may escape detection.

Definition

If an island contains exactly one clone, it is a *singleton island* or a trivial contig; otherwise, it is a non-trivial *contig* containing at least two clones.

Definition

An *ocean* is a region of the genome between two neighboring islands.

1. *How are the islands distributed?* ... Let us introduce the indicator variable $S_{i,j}$ denoting the event that the clone i starts at the position j of the genome:

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

2. Let

$$V_a = \sum_{i=1}^{N} \sum_{j=a-L+T}^{a-1} S_{i,j}$$

be the random variable denoting the number of clones covering the position a from the left by the amount no larger than $L - T$.

3. Thus every clone starting at position a will be detected to overlap with V_a clones "from the left," as they will each overlap in T base pairs or more with a clone starting at a.

1. Hence,

$$\begin{aligned}
\mathbb{E}\left[\binom{V_a}{k}\right] &= \binom{(L-T)N}{k}\left(\frac{1}{G}\right)^k \\
&\approx \left[\frac{(L-T)N^k}{G}\right]/k! \\
&= \frac{(c-(T/L)c)^k}{k!} \\
&= \frac{(c\sigma)^k}{k!}
\end{aligned}$$

2. Hence, by Brun's sieve, we have:

$$\mathbb{P}r[V_a = k] = \mathbb{E}\left[\mathbb{I}_{V_a=k}\right] = e^{-c\sigma}\frac{(c\sigma)^k}{k!}.$$

In particular:

$$\mathbb{P}r[V_a = 0] = e^{-c\sigma}, \quad \text{and} \quad \mathbb{P}r[V_a \neq 0] = 1 - e^{-c\sigma}.$$

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 $\times$ coverage clone library for human (as claimed by Celera), then we need to use a $\theta \leq 0.474$.

- The probability that the ith 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

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

- 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.

Outline

- 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 show in the next lecture (with a more detailed analysis) that the above expression is correct for all values of θ.

[End of Lecture #3]