Shotgun sequencing

Rahul Roy

Indian Statistical Institute, New Delhi

Introduction

A single strand of the human DNA has nearly 3.3 billion nucleotides.

It is not possible for a single laboratory to sequence this genome, i.e. identify each of the nucleotides comprising this genome.

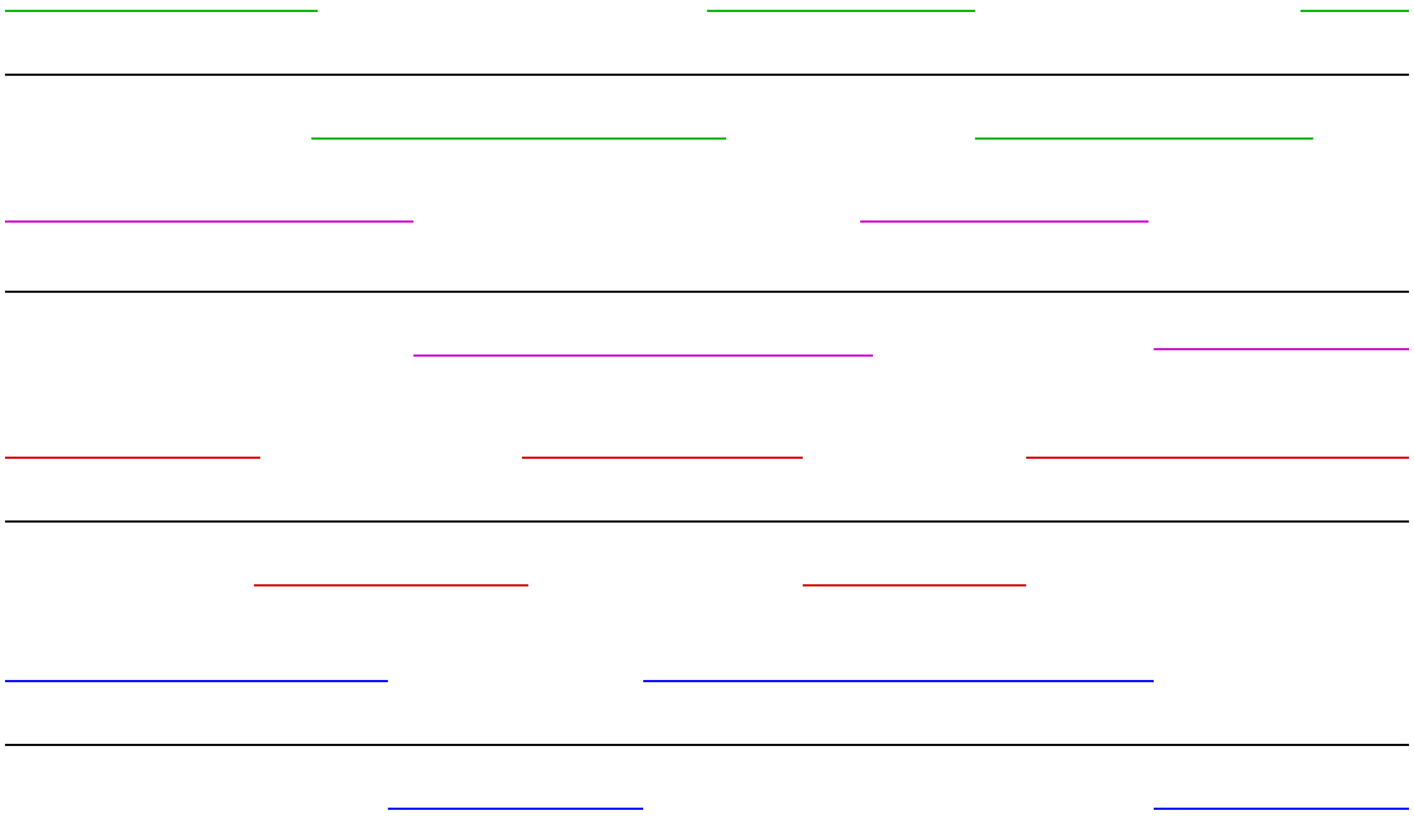
Shotgun sequencing is a method to identify the nucleotides forming the sequence.

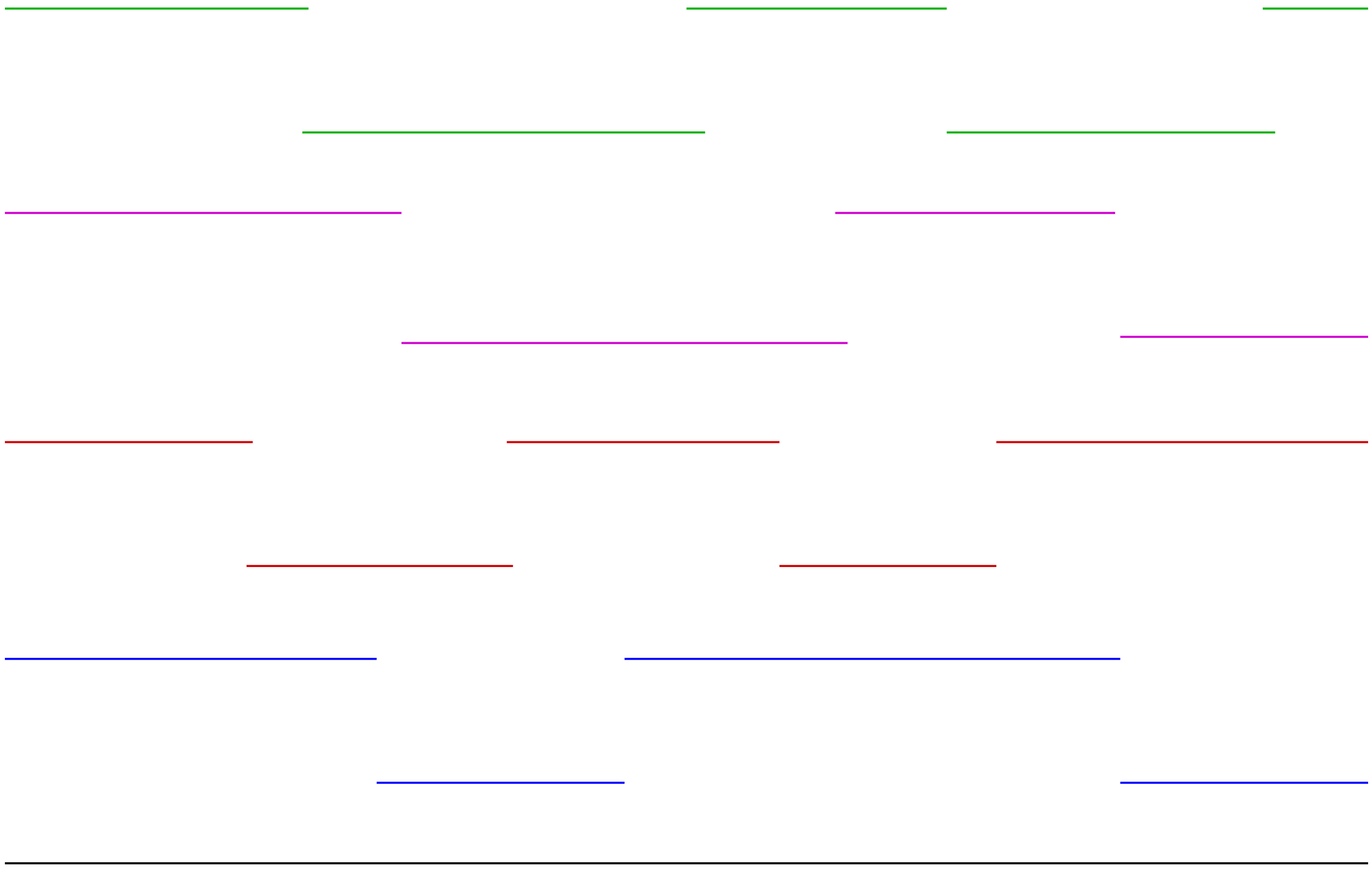
In this talk we outline the methods from probability involved in shotgun sequencing and then explain a mathematics result which is an offshoot of this.

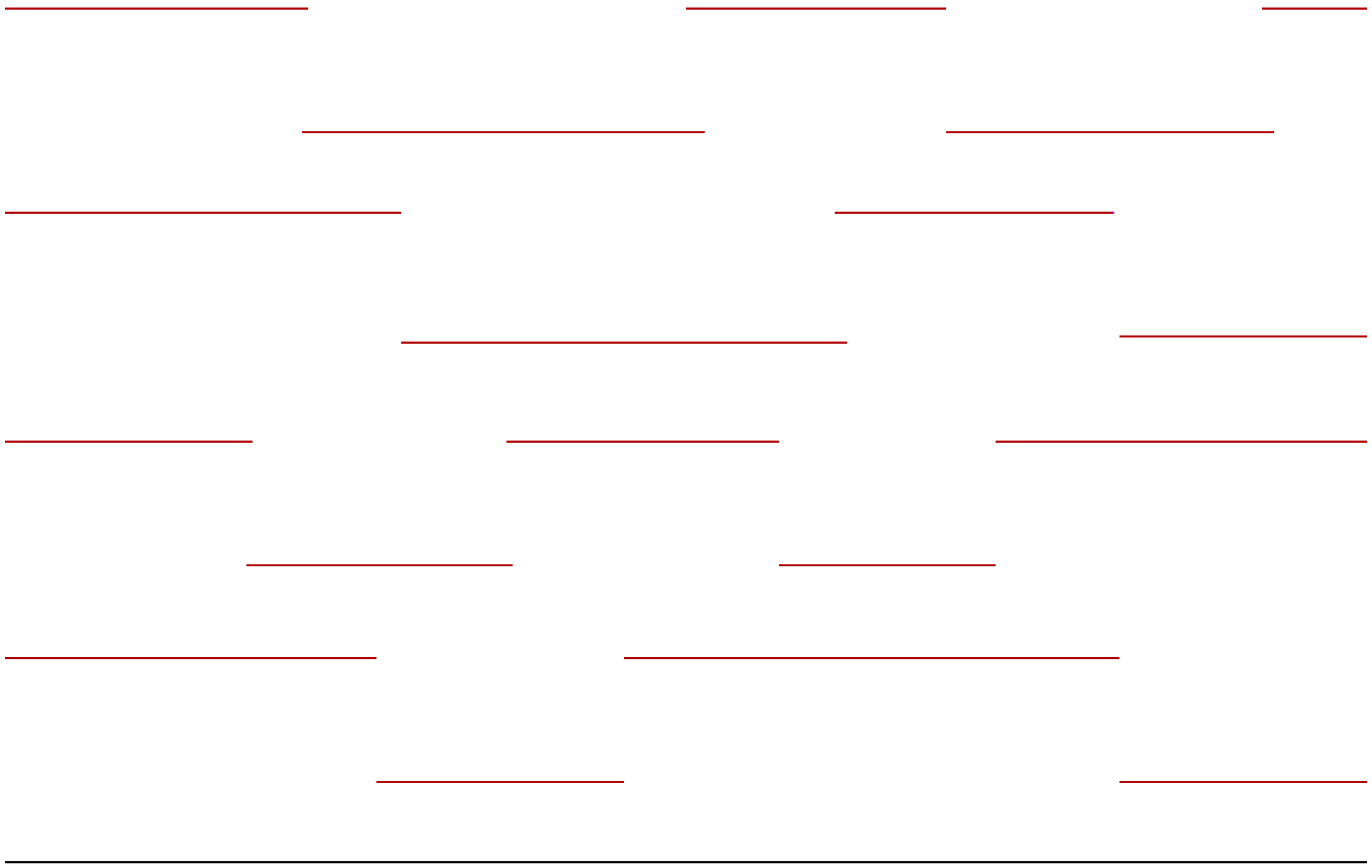
References

1. Arratia, Lander, Tavaré, Waterman, **Genomics** (1991)
2. Schbath, **J. Comp. Biology** (1997)
3. Schbath, Bossard, Tavaré, **J. Comp. Biology** (2000)
4. Athreya, Roy, Sarkar, **Adv. Appl. Probab.** (2004)

We begin with a library of clones.
Clones are segments of the genome and are typically much smaller in size.
In the human genome project these clones were around 100,000 nucleotides in length.
These clones are obtain by biochemical means and are obtained from multiple copies of the gene, and through different biochemical means so as to ensure that for each clone there are other clones which overlap.







After identifying the nucleotides constituting each clone, the task is to stitch the clones together to obtain the entire genome. The problems during stitching the clones are as follows:-

(i) since the genome is much larger than the clones and also because of the biochemical means involved in obtaining the clones, it is not possible to keep a track of the order of appearance of the clones in the genome,

(ii) the amount of overlap required has to be optimized because a short overlap increases the chances of error, while a long overlap demands more clones to cover the sequence and as such more effort.

These problems may be slightly alleviated by the presence of anchors in the genome.

Anchors are short unique stretches of nucleotides which are either inserted in the genome, or whose presence is known in the genome.

These anchors perform the role of markers in the genome; each marker being unique in its nucleotide composition and being randomly located in the genome.

Overlapping clones are stitched into islands and the maximal such connected regions forming contigs.

The aim of this sequencing scheme is to obtain contigs which cover as much as possible of the genome, leaving as little gaps (oceans) as possible.

Mathematical model

For a mathematical modelling of the sequencing method we start with some assumptions. The most fundamental assumptions pertaining to the genome itself are

(i) the genome is a continuous, rather than a discrete system,

(ii) the genome is a bounded segment of a doubly infinite random string of nucleotides, the random process producing this doubly infinite string of nucleotides is stationary and also ergodic.

The genome will be taken to be, without loss of generality, the interval $[0, G]$.

It is also assumed that

(iii) the clones are segments of length $L_1, L_2, \ldots$, where $L_1, L_2, \ldots$ is an i.i.d. bounded sequence of random variables, each with a probability density function $f(l)$, and independent of the random process producing the genome.

Without loss of any generality we take $E(L_1) = 1$.

The centre points of the clones are uniformly distributed in the segment $[0, G]$; there being a random number N of clones whose centre points lie in $[0, G]$. This is guaranteed by the assumption

(iv) the centre points of the clones are distributed according to a Poisson point process of density a on $\mathbb{R}$ and independent of all other processes described above,

with $a = E(N)/G$, where N is the random number of clones centred in the genome $[0, G]$.
This quantity a is called the coverage of the sequencing method and denotes the factor by which the expected total length of the clones exceeds the length of the genome.

Finally, the random number M of anchors is distributed uniformly in the genome and this is obtained by the assumption that

(v) the anchors are distributed according to a Poisson point process of density b on $\mathbb{R}$ and independent of all other processes as yet,

where $b = E(M)/G$; M being the random number of clones centred in the genome $[0, G]$.

Let $\{X_i : i \in \mathbb{Z}\}$ be a labelling of the right end points of the clones – the labeliing being such that

$$\cdots < X_{-1} < X_0 \leq 0 < X_1 < \cdots < X_{N'} < G \leq X_{N'+1} < \cdots ,$$

N' being the random number of clones whose right end points fall within the genome $[0, G]$.
In this notation a contig is an interval $[X_i - L_i, X_k]$ s.t.

(a) for any $x \in [X_i - L_i, X_k]$ there exists j such that $x \in [X_j - L_j, X_j]$

(b) for every $\epsilon > 0$, there exist $y \in [X_i - L_i - \epsilon, X_i - L_i]$ and $z \in [X_k, X_k + \epsilon]$ such that $x, y \notin \cup_t [X_t - L_t, X_t]$

i.e. a connected component of the region
$C = \cup_{i \in \mathbb{Z}} [X_i - L_i, X_i]$

N' has a Poisson distribution with mean aG, i.e.

$$P(N' = n) = \frac{e^{-aG}(aG)^n}{n!} \text{ for } n \geq 0.$$

Also from the theory of Poisson processes we know that the differences $\{X_i - X_{i-1} : i \in \mathbb{Z}\}$ are i.i.d. exponential random variables with mean $1/a$, i.e. $X_i - X_{i-1}$ has a probability density function

$$f(x) = \begin{cases} e^{-ax} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

These differences are exactly the inter-arrival times of probability theory.

By the independence of the anchor process and clone process, the anchored clones and the unanchored clones each form thinned processes of the original Poisson process of the clones – the anchored clones forming a Poisson process of density

$aP(\{\text{a clone of length } L \text{ has at least one anchor}\})$

and the unanchored clones forming another Poisson process of density

$aP(\{\text{a clone of length } L \text{ has no anchors}\})$.

These processes are independent of each other, and,

$$\begin{aligned} P\{\text{a clone of length } L \text{ has no anchors}\} &= \int_0^\infty e^{-bl} f(l)\,dl \\ &= E(e^{-bL}) \\ &= q_1 \text{ (say)}, \end{aligned}$$

where $f(l)$ is the probability density function of L.
So if N_u and N_a denote the number of unanchored and anchored clones whose right end-points fall on te genome $[0, G]$, then

$$N_u \sim \text{Poisson}(aGq_1) \text{ and } N_a \sim \text{Poisson}(aG(1 - q_1))$$

Also, as $G \to \infty$, almost surely we have

$$\frac{N_u}{G} \longrightarrow aq_1$$

and

$$\frac{N_a}{G} \longrightarrow a(1 - q_1).$$

Anchored Island

Let the anchor process be labelled as

$$\cdots < Y_{-1} < Y_0 \leq 0 < Y_1 < \cdots < Y_M < G \leq Y_{M+1} < \cdots ,$$

An anchored island or contig is a connected region formed by consecutively intersecting clones such that there is an anchor in each of these intersecting regions, and also the island is maximal with respect to these properties.

Formally an anchored island is a connected region formed by clones with right-end points $X_{i_1}, \ldots, X_{i_k}$ and of lengths $L_{i_1}, \ldots, L_{i_k}$respectively for some k, such that

(i) $X_{i_1} < \cdots < X_{i_k}$,

(ii) for every j there exists a $l \neq j$ and an anchor Y_{m_j} such that $Y_{m_j} \in [X_{i_j} - L_{i_j}, X_{i_j}] \cap [X_{i_l} - L_{i_l}, X_{i_l}]$,

(iii) the region $[X_{i_1} - L_{i_1}, X_{i_k}]$ is not (strictly) contained in an island containing clones satisfying (i) and (ii) above.

Let

$$\cdots < C_{-1} < C_0 \leq 0 < C_1 < \cdots < C_K < G \leq C_{K+1} < \cdots$$

be a labelling of the right end-points of anchored contigs. K denoting the random number of these right end-points falling in $[0, G]$.
Note that $\{C_i\}$ is also a Poisson process thinned out of the process $\{X_i\}$. The process $\{C_i\}$ has density ap_1 where

$$p_1 := P\{X_j \text{ is the right end-point of an anchored contig}\}$$

To understand p_1 let 0 be the right end-point of an anchored contig.

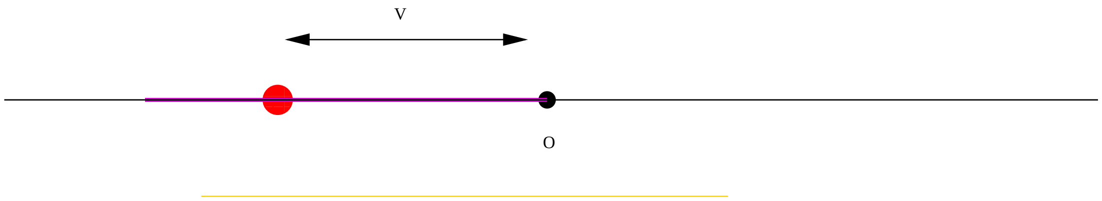
Note we may do this because for a Poisson process $\{\xi_i\}$, the distribution of the process given $\xi_0 = 0$ is the same as the unconditional distribution.

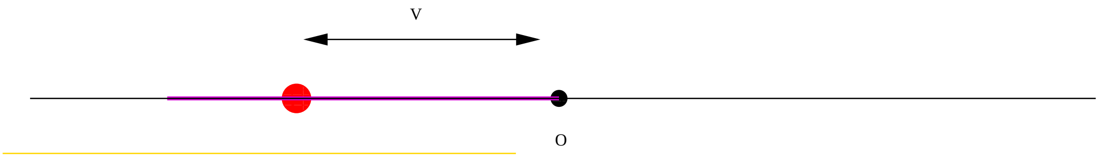
For 0 to be the right end-point of an anchored contig we must have

i) there is a clone with an anchor of length L with right end-point 0,

ii) For some $0 < V < L$ there is an anchor located at $-V$ which is the anchor left nearest to 0,

iii) any clone with left end-point $\leq -V$ must have its right end-point ≤ 0.


V
O


V
O

Thus
$p_1 = P\{$ (i), (ii) and (iii) hold $|$ 0 is the right end-point of a clone$\}$

Let $R(l) = P(L > l)$ and $J(x) = \exp\{-a \int_x^\infty R(l)dl\}$.

Note that
$J(x) = P\{$ 0 and x are not covered by a common clone$\}$.
With this notation

$$\begin{aligned} p_1 &= E(I_{V \leq L} J(V)) \\ &= \int_0^\infty be^{-bu} J(u)R(u)du. \end{aligned}$$

Proposition 1

So we have proved

Proposition 1

a)
$P($a clone contains no anchors$) = \int_0^\infty e^{-bl} f(l) dl$
and
$N_u \sim$ Poi$(Ga \int_0^\infty e^{-bl} f(l) dl)$.

b)
$P($a clone is the rightmost clone of an anchored contig$) = \int_0^\infty be^{-bu} J(u) R(u) du$
and
$K \sim$ Poi$(Ga \int_0^\infty be^{-bu} J(u) R(u) du)$.

Proposition 2

Now we show

Proposition 2

The expected number of anchored clones in an anchored contig is $\frac{1-q_1}{p_1}$.

Proof:
Consider the K anchored islands in the genome $[0, G]$. Let M_j be the number of anchored clones in the jth anchored island counted from the left.

Unfortunately, $M_1, M_2, \ldots, M_K$ is not stationary because the labelling destroys staionarity. In particular, because of the waiting time paradox, the first anchored contig which straddles 0 has length M_1 which is larger than *typical*.

Let

$$\bar{M}_G = \frac{1}{K}\sum_{j=1}^{K} M_j \text{ and } M'_G = \sum_{j=1}^{K} M_j - N_a$$

Note that N_a is the number of anchored clones which has right end-points in $[0, G]$, and each of these clones are counted in $\sum_{j=1}^{K} M_j$, so the difference $\sum_{j=1}^{K} M_j - N_a$ is at most the number of clones in the anchored contigs covering 0 or G.

So

$$\sum_{j=1}^{K} M_j - N_a \leq M_1 + M_K (= M'').$$

Now $E(M'') < \infty$, so

$$\frac{M'_G}{G} \rightarrow 0 \text{ a.s. as } G \rightarrow \infty$$

Thus, almost surely,

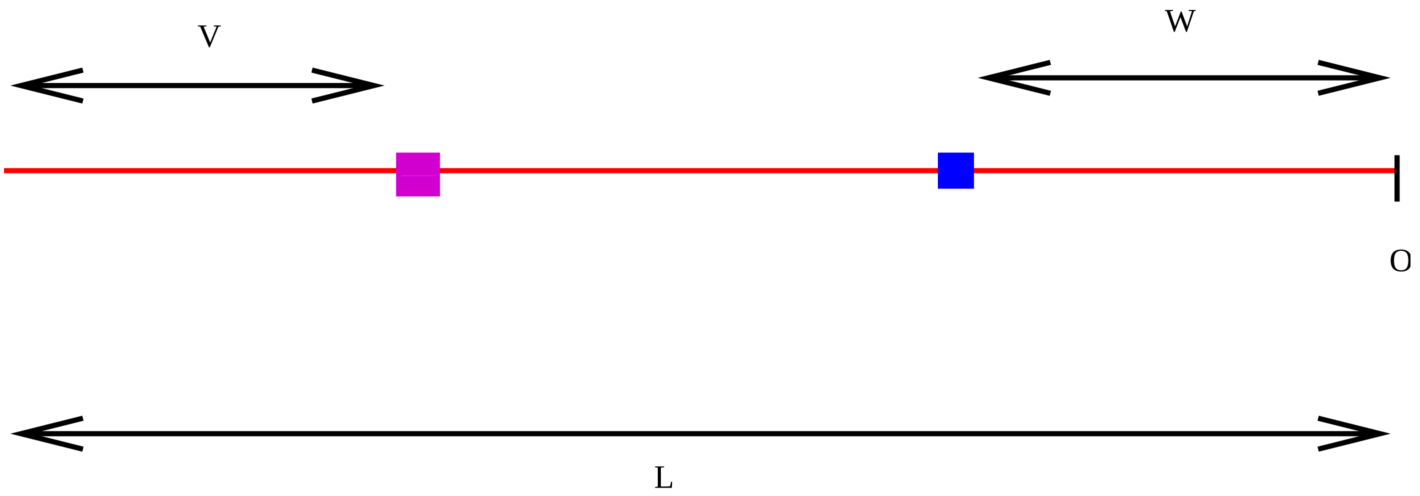
$$\begin{aligned}
\bar{M}_G &= \frac{G}{K}\frac{1}{G}\sum_{j=1}^{K} M_j \\
&= \frac{G}{K}\left[\frac{N_a}{G} + \frac{M'_G}{G}\right] \\
&\to \frac{1}{ap_1}a(1-q_1) = \frac{1-q_1}{p_1} \text{ as } G \to \infty
\end{aligned}$$

because

$$\frac{K}{G} \to ap_1, \frac{N_a}{G} \to a(1-q_1), \frac{M'_G}{G} \to 0.$$

Singleton anchored clone

Now consider the event E that a given clone is a singleton anchored contig. Suppose this clone has length L and the right end-point is 0. Also suppose the two extreme anchors on this clone are located as in the figure below.



So $V + W = L$ if there is only one anchor on the clone, otherwise $V + W < L$.

Thus

$$\begin{aligned} & P(E|V,W \text{ and } V+W<L) \\ &\quad = P\{ \text{ no clones starting to the left of } -L \\ &\qquad \text{cover the anchor at } -L+V\text{, and no clones} \\ &\qquad \text{starting in } (-L,-W) \text{ cover } 0\} \\ &\quad = \frac{J(V)J(W)}{J(L)}. \end{aligned}$$

And similarly,

$$P(E|V,W \text{ and } V+W=L) = \frac{J(V)J(L-V)}{J(L)}.$$

To remove the conditioning note that if V' and W' are two exponential random variables with parameter b, then

$$(V,W) \overset{\mathcal{L}}{=} (V',W')1_{\{V'+W'<L\}} + (V',L-V')1_{\{V'<L,V'+W'\geq L\}}.$$

Thus

$$\begin{aligned} P(E) &= E\left(1_{\{V'<L\}}\frac{J(V')J(\min\{W', L-V'\})}{J(L)}\right) \\ &= \int_0^\infty \int_0^l \int_0^{l-v} b^2 e^{-b(u+v)}\frac{J(u)J(v)}{J(l)} f(l)dudvdl \\ &\quad + \int_0^\infty \int_0^l be^{-bl}\frac{J(u)J(l-u)}{J(l)} f(l)dudl \\ &= p_2(\text{say}). \end{aligned}$$

Moreover the right-end points of clones which are singleton anchored contigs form a Poisson point process. So we have

Proposition 3

$$P\{\text{a clone is a singleton anchored contig}\} = p_2$$

and the expected number of such singleton anchored contigs in the genome is Gap_2.

Length of an anchored contig

Now we find the expected length of an anchored contig.
Let S_j denote the length of the jth anchored contig from the origin 0.
Let

$$\bar{S}_G = \frac{1}{K}\sum_{j=1}^{K} S_j.$$

By ergodicity we have

$$E(S) = \lim_{G\to\infty} \bar{S}_G \text{ almost surely,}$$

where S is the random length of an anchored contig.

Observe that $\sum_{j=1}^{K} S_j$ measures the length of the region in the genome covered by anchored contigs with the caveat

(a) the contigs which straddle either 0 or G contribute regions to $\sum_{j=1}^{K} S_j$ which lie *outside* $[0, G]$

(b) the region covered by two distinct anchored contigs is double counted in $\sum_{j=1}^{K} S_j$.

By ergodicity we know that $\lim_{G\to\infty} \frac{1}{G} \sum_{j=1}^{K} S_j$ gives the density of the points covered by the contigs – a density which respects double counting. Thus

$$\frac{1}{G} \sum_{j=1}^{K} S_j \longrightarrow r_1 + 2r_2 \text{ a.s. } G \to \infty.$$

where
$r_i = P($ a given point $x \in [0, G]$ is covered by i distinct anchored contigs$)$.

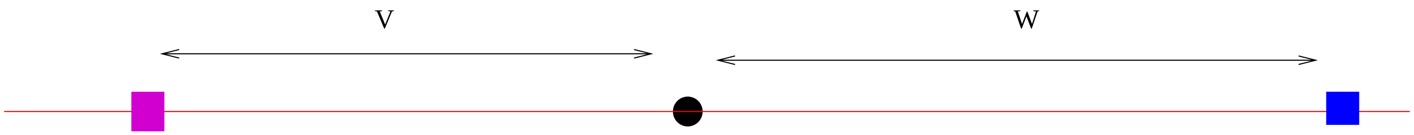
To obtain r_1 and r_2 we note that

$$r_0 + r_1 + r_2 = 1$$

so we need only find r_0 and r_2.
Let

$$E_0 = \{t \text{ is not covered by an anchored contig}\}$$

so that $r_0 = P(E_0)$.



For E_0 to occur we must have

(a) any clone starting to the left of $t - V$ must end before t

(b) any clone starting in $(t - V, t)$ must end before $t + W$.

Thus

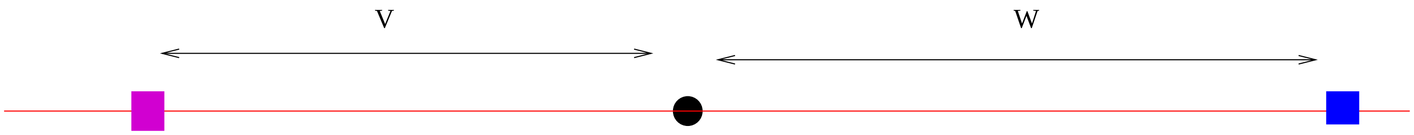
$$\begin{aligned} P(E_0|V,W) &= \exp\left(-a\int_V^\infty R(r)dr - a\int_W^{V+W} R(r)dr\right) \\ &= \frac{J(V)J(W)}{J(V+W)}. \end{aligned}$$

So we have

$$
\begin{aligned}
r_0 &= E\left[P(E_0|V,W)\right] \\
&= E\left(\frac{J(V)J(W)}{J(V+W)}\right) \\
&= \int_0^\infty \int_0^\infty b^2 e^{-b(u+v)} \frac{J(u)J(v)}{J(u+v)} du dv
\end{aligned}
$$

Now let

$$E_2 = \{t \text{ is covered by two distinct anchored contigs}\}$$



For E_2 to occur we must have

(a) at least one clone starting to the left of $t - V$ ends in $(t, t + W)$,

(b) no clone starting to the left of $t - V$ ends after $t + W$,

(c) at least one clone starts in $(t - V, t)$ and ends after $t + W$.

Thus

$$P(E_2|V,W) = \left(1 - \frac{J(V)}{J(V+W)}\right) J(V+W) \left(1 - \frac{J(W)}{J(V+W)}\right).$$

Thus, after simplification, we have

$$\begin{aligned} r_1 + 2r_2 &= 1 - r_0 + r_2 \\ &= 1 + EJ(V+W) - 2EJ(V). \end{aligned}$$

Proposition 4

Hence, we have our result

$$\begin{aligned} E(S) &= \lim_{G\to\infty} \frac{G}{K}\frac{1}{G}\sum_{j=1}^{K} S_j \\ &= \frac{1}{ap_1}(1 + EJ(V+W) - 2EJ(V)) \\ &= \frac{1}{ap_1}\left(1 + \int_0^\infty (b^2u - 2b)e^{-bu}J(u)du\right). \end{aligned}$$

We also have for free the result that the expected proportion of the genome not covered by an anchored contig is

$$r_0 = \int_0^\infty \int_0^\infty b^2 e^{-b(u+v)} \frac{J(u)J(v)}{J(u+v)} dudv.$$

Number of anchors in an anchored contig

First note that

$$P(\text{an anchor is not covered by a clone}) = J(0) = e^{-a}.$$

Thus the density of anchors which are covered by clones is $b(1 - e^{-a})$.
By ergodicity, denoting by R the number of anchors in $[0, G]$ covered by clones we have

$$\frac{R}{G} \longrightarrow b(1 - e^{-a}) \text{ a.s. as } G \to \infty.$$

Hence, writing H_j as the number of anchors in the jth anchored contig and H as the random number of of anchors in an anchored contig,

Proposition 5

$$
\begin{aligned}
E(H) &= \lim_{G\to\infty} \frac{1}{K}\sum_{j=1}^{K} H_j \text{ a.s.} \\
&= \lim_{G\to\infty} \frac{G}{K}\frac{R}{G} \\
&= \frac{b(1-e^{-a})}{ap_1}.
\end{aligned}
$$

Besides the anchored islands, there are also oceans or gaps, i.e. the region between the islands. An ocean may be actual in the sense that there is no clone unanchored situated there or apparent if there are clones unanchored and hence undetected in the ocean.

Proposition 6

For any $x \geq 0$, the probability that an anchored island is followed by an actual ocean of length at least x is $e^{-a(x+1)}\dfrac{1-q_1}{p_1}$.

Thus taking $x = 0$ we have the probability that an anchored island is followed by an actual ocean rather than an undetected overlap.

Proof: Given a clone of length L whose right-end point is situated at the origin 0 and V the distance of the left-nearest anchor,

the conditional probability that this clone is
a) anchored,
b) the end of its island, and
c) followed by an actual ocean of length at least x

is the probability that

a) $V < L$,
b) all clones that start (i.e. left-end point) to the left of 0 also end to the left of 0, and
c) no clones have their left-end point in the interval $[0, x]$.

This conditional probability is

$$P(1_{\{V<L\}})J(0)e^{-ax} = (1-q_1)e^{-a(x+1)}.$$

Thus recalling that p_1 also denotes the probability that a given clone is the end of an anchored contig, we get our Proposition.

Remark 1 In case $L \sim$ Unif$[1-s, 1+s]$ then

$$J(x) = \begin{cases} \exp\left(-a(1-x)\right) \text{ for } 0 \le x \le 1-s \\ \exp\left(-a\frac{(1+s-x)^2}{4s}\right) \text{ for } 1-s \le x \le 1+x \\ 1 \text{ for } x > 1+s. \end{cases}$$

Remark 2 If $L \sim$ Exp(1) then $J(x) = \exp(-ae^{-x})$.

Mathematical Curiosity

Proposition: The expected number of anchored islands is always larger for the case of clones having constant length than for the case of clones having variable length, but the same mean.

Proof: Observe that

$$\begin{aligned} J(x) &= \exp\left(-a\int_x^\infty P(L > l)dl\right) \text{ for } x > 0 \\ &= \exp\left(-aE(\max\{L - x, 0\}\right). \end{aligned}$$

Since the function $\max\{x, 0\}$ is convex, by Jensen's inequality we have

$$
\begin{aligned}
J(x) &\leq \exp\left(-a(\max\{E(L-x),0\}\right) && (1)\\
&= J_{\text{const}}(x), && (2)
\end{aligned}
$$

where

$$
J_{\text{const}}(x) = \begin{cases} \exp(-a) \text{ for } 0 \leq x \leq 1 \\ 1 \text{ for } x > 1, \end{cases}
$$

i.e. J is a constant for $L \equiv 1$.

From Proposition 1 we have

$$\begin{aligned} p_1 &= \int_0^\infty be^{-bu} J(u) R(u) du \\ &= \frac{1}{a} \int_0^\infty be^{-bu} J'(u) du. \end{aligned}$$

Integrating by parts and using the fact that $J(0) = e^{-a}$, we have

$$ap_1 = -be^{-a} + \int_0^\infty b^2 e^{-bu} J(u) du.$$

Now from the inequality (1) we have

$$ap_1(L) \leq ap_1(\text{constant}).$$

A negative result

We now think of the genome as a half integer line $\mathbf{N}$.
Let $X_1, X_2, \ldots$ be a Markov chain taking values 0 or 1. Also let $L_1, L_2, \ldots$ be i.i.d. random variables having the same distribution as a random variable L.
Let

$$C = \cup_{i \geq 1} I_{\{X_i = 1\}} [i, i + L_i].$$

Here the analogy is that a clone of length L_i is cut at i and C is the region covered by the contig in the half-line.

If $p_{ij} = P(X_{n+1} = j \mid X_n = i)$, $i, j = 0$ or 1, denote the transition probabilities of the Markov chain then we have

Proposition

Suppose $0 < p_{00}, p_{10} < 1$.

(a) If $a := \liminf_{j\to\infty} jP(L > j) > 1$, then $P\{[t,\infty) \subseteq C$ for some $t\} = 1$ whenever $\frac{p_{01}}{p_{10}+p_{01}} > 1/a$.

(b) If $A := \limsup_{j\to\infty} jP(L > j) < \infty$, then $P\{[t,\infty) \subseteq C$ for some $t\} = 0$ whenever $\frac{p_{01}}{p_{10}+p_{01}} < 1/A$.