DOUBLE-END CLONE SEQUENCE ASSEMBLY: SOME STATISTICAL ISSUES

Xiaoman Li[1], Michael S.Waterman[1,2]
[1]Department of Mathematics, University of Southern California,
1024 West 36th Place, Los Angeles, CA90089-1113, USA
[2]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850

ABSTRACT

Motivation: Recently, in many genome sequencing projects, people have used double-end clones. This paper is concerned with predicting genome coverage in these projects. The Lander-Waterman formula can only address the statistical properties for the assembly projects using clones without "mate-pairs". There is a need to extend the Lander-Waterman formula to cover double-end genome sequencing.
Results: We improve prior results and calculate the average number and length of scaffolds, islands, and gaps; estimate the average number of islands in a scaffold; and so on. In addition, we estimate the distribution of the gap size between adjacent islands.
Contact: xiaomanl@usc.edu. Tel: (213)740-0778. Fax: (213)740-2424.
keywords: DNA sequencing, double-end clone, scaffold, island, Poisson process, Lander-Waterman formula.

1. Introduction

There are so many repeats in the human and other genomes that scientists have used the double-end clone strategy to "walk" through them (Venter et al., 1998; Webber and Myers, 1997). This makes it possible to sequence many genomes but many complex problems in the assembly process have arisen. One is to understand the stochastic process of the assembly by using double-end clones.

In 1995 Port et al. addressed some statistical problems about double-end clone mapping and sequencing strategies. In that paper, the authors attempted to extend the Lander-Waterman formula; i.e., they wanted to estimate the average length and the average number of scaffolds, of islands, of gaps, and the average number of islands and gaps in a scaffold, and so on, for the double-end clone strategy. (Scaffolds are clones connected via mate-pairs or double ends). That paper established some results and posed the problem rigorously. But neither their greedy island method nor their block-island method can give a good description of scaffolds, such as the average number and length of scaffolds, even for length-fixed clones with length-fixed ends. Exactly because of the complexity of the problem, people such as Roach et al. (Roach et al., 1995; Siegel et al., 2000) used simulations to get some idea about the result of assembly by using double-end clones.

In 1999, Ru-Fang Yeh improved Port et al.'s results very much in her thesis by assuming the lengths of clones and ends are random variables. First, she made

a correction for the Lander-Waterman formula by adding a term to describe the boundary effects which are important in high-coverage cases (i.e. the previous formula estimates the island number as 0 when the coverage is infinite). Then she gave a very good approximation to the average number of the scaffolds, although there are some errors in the arguments. In addition, she gave some other formulae to estimate other quantities mentioned above.

As did Port et al, Yeh advanced the state of knowledge. However, only the average number of scaffolds is reliable. This will be discussed in detail later. In this paper, we will study the estimation of scaffolds, islands and gaps. Our study is intended for genome sequencing projects.

2. Estimation for double-end clone strategy

2.1. **Basic Concepts and Notations.** By double-end clones, we mean clones where we sequence the two ends, about 500bp at each end, say, instead of sequencing just one fragment. The sizes of clones Celera used in their human genome project are approximately 2kb, 10kb and 50kb (Venter et al., 2001). By the double-end clone strategy, we mean the method in which a double-end clone library is constructed, and random clones are chosen for their clone ends to be sequenced. Then scientists try to assembly those clones to cover the whole genome. Here and in the following, we will use ends to mean the leftmost and the rightmost parts of the clones (in reference to a genome sequence) that have been sequenced. By scaffolds, we mean a set of clones, any two of which are connected by direct or indirect overlaps of ends. Of course, there may be many gaps in a scaffold. By islands, we mean a set of ends that are connected with each other by direct or indirect overlaps (not using double-end connections). Thus, there is no gap in an island. By gaps, we mean those regions that are between two adjacent islands that are not covered by the clone ends.

Now, we have the following notation:

G= genome length;
N = number of clones;
E = the average length of ends;
L = the average length of clone;
T = length of common substrings needed to detect overlap;
$\theta = \frac{T}{L}$; $\lambda = \frac{N}{G-L+1}$; coverage$\stackrel{\triangle}{=} \frac{2NE}{G}$;
L_t is the clone starting at t or its length, $t \in (0, G)$;
$F(\bullet)$ = the distribution function of the length of clones;
E_{tl} is the length of the left end of the clone L_t, $t \in (0, G)$;
E_{tr} is the length of the right end of the clone L_t, $t \in (0, G)$;
$G(\bullet)$ = the distribution function of the length of ends of clones.

2.2. **Previous Work.** In Port et al.'s paper (Port et al., 1995), they studied length-fixed double-end clones with length-fixed ends by two methods. The following is one of their approximations.

$$\begin{aligned} p(t) \triangleq & Pr(\text{a scaffold ends at t}|\text{a clone ends at t}) \\ \approx & Pr(\text{no other clone ends after t with left end overlapping with} \\ & \text{either end of the clone ending at t}|\text{a clone ends at t }) \end{aligned}$$

The approximation defined above is easy to calculate since we model the leftmost positions of clones as a Poisson process (Waterman, 1995). But the result of $p(t) = e^{-3\lambda E}$ in Port et al. is too imprecise to be a good estimator.

Keeping this definition in mind, and assuming the length of clones and ends are random variables with known distribution, Yeh (Yeh, 1999) defined

$$\begin{aligned} p(t) \triangleq & Pr(\text{a scaffold starts at t}|\text{a clone starts at t}) \\ \approx & Pr(A|\text{a clone starts at t}) + Pr(A^cB|\text{a clone starts at t}) \end{aligned}$$

Where

$$\begin{aligned} A = & \{\text{no clone starts before t and ends after t}\}; \\ B = & \{\text{both the clone in } A^c \text{ and the clone } L_t \text{ do not overlap with} \\ & \text{the end of any other clone beginning after t and before } t + L_t - E_{tr}\}; \end{aligned}$$

and of course A^c is the complement of the set A. Note that the two equations for $p(t)$ above describe the same thing although one considers the end of a scaffold while the other considers the start of a scaffold.

Although the approximation by Yeh avoids the most difficult parts in calculating $p(t)$ exactly[1], it is good enough for application. Her calculation resulted in the following formula:

$$p(t) \approx e^{-\lambda L} + e^{-5\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-3E)})(2 - e^{-\beta\lambda E}) \tag{1}$$

where

$$\begin{aligned} & Pr(A|\text{a clone starts at t}) \approx e^{-\lambda L}, \\ & Pr(A^c|\text{ a clone starts at t}) \approx e^{-3\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-3E)}), \\ & Pr(B|A^c \text{ and a clone starts at t}) \approx e^{-2\lambda E}(2 - e^{-\beta\lambda E}), \end{aligned}$$

and β and α are two constants determined by the length distribution of clones and ends (Yeh, 1999).

Then, by adding a correction item for the left boundary 0, Yeh obtained

$$\begin{aligned} & E(\text{number of scaffolds}) \\ \approx & 1 - e^{-\lambda L} + Np(t) \\ \approx & 1 - e^{-\lambda L} + N[e^{-\lambda L} + e^{-5\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-3E)})(2 - e^{-\beta\lambda E})]. \end{aligned} \tag{2}$$

Equation (2) is such a good estimation for the average number of scaffolds in that it agree with our simulation results[2] very well when coverage is larger than 2;

[1] See Figure 2 and the paragraph after equation 2.

[2] Our simulation parameters are as follows. The genome length is 80kb. There are two clone lengths, one of that is 5kb, the other 2kb. Each clone length accounts for 50% of the total number

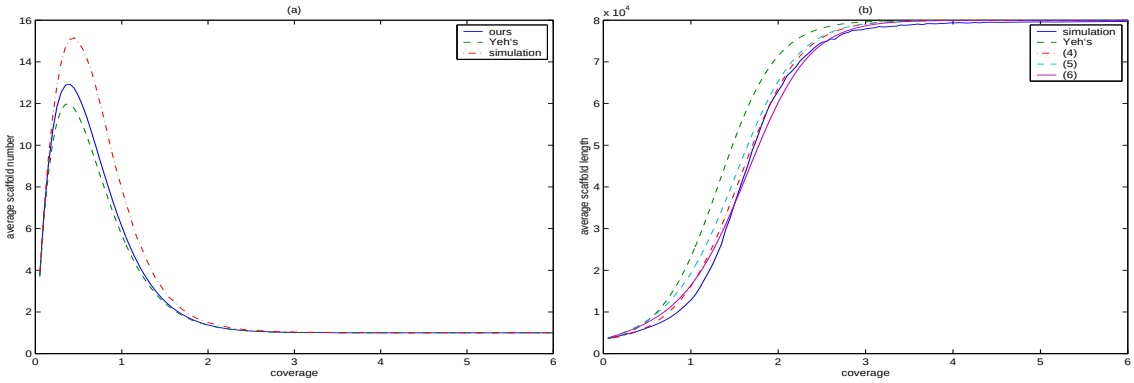

FIGURE 1. (a)Average number of scaffolds as a function of coverage.(b)Average length of scaffolds as a function of coverage.

see Figure 1(a). But there are some problems with the calculation. First, when considering event B, only those clones with left end overlapping with the ends of L_t or the right end of the clone in A^c, such as L_{s1} in Figure 2(c), are included. So the estimate neglects those clones such as L_{s2} in Figure 2(c). Second, when calculating $Pr(B|A^c$and a clone starts at t), Yeh (Yeh,1999) only considered those clones starting before t and ending before $t + L_t$, such as L_{s1} in Figure 2(a), and neglected those starting before t and ending after $t + L_t$, such as L_{s2} in Figure 2(a). Third, there may be many clones in A^c, such as L_{s1}, L_{s2}, L_{s3} in Figure 2(b), and all of them should be included. But Yeh's argument is based on the assumption that there is only one clone in A^c.

Yeh also tried to estimate the average length of scaffolds, the average length of gaps and the average number of islands in a scaffold. But those estimations are not of good quality. We will show the comparison of Yeh's and ours in detail later.

2.3. **Our Estimates.**

2.3.1. *Average Scaffold Number.* We define the following sets.

$$S1 \stackrel{\triangle}{=} \{\text{clone } L_s | s < t < t + E_{tl} < s + L_s - E_{sr} < s + L_s < t + L_t - E_{tr}\}$$

$$S2 \stackrel{\triangle}{=} \{\text{clone } L_s | s < t < t + L_t < s + L_s - E_{sr}\}$$

As we pointed out above, the clones in A^c belongs to $S1 \bigcup S2$. Yeh (Yeh,1999) only considered the clones in $S1$. We now consider the clones in $S2$.

For given L_t, it is possible that $|S1| > 1$ or $|S2| > 1$. Although we know it is wrong to think there is only one clone in $S1$ as Yeh (Yeh,1999) did, we will follow that method of calculation because it is difficult to analyze what happens between those clones and L_t. For instance, in Figure 2(b), there are only three clones. We have to use triple integral to calculate $Pr(A^c \bigcap B \bigcap S1|$a clone starts at t). It is difficult

of clones. The length of the ends is fixed as 500bp. We also assume T to be 25bp for detecting overlap. For each point on the graph, we simulated 3000 times and got the average.

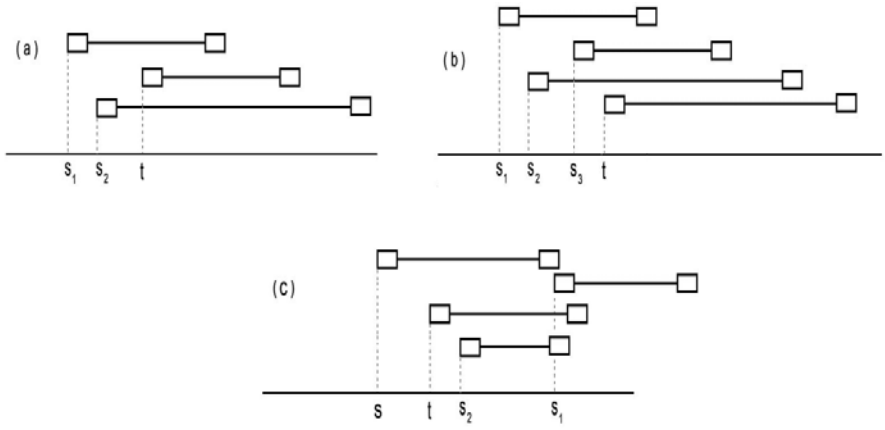

FIGURE 2

enough when there are more than three clones. So we will assume there is only one clone in $A^c \bigcap B \bigcap S2$ when we calculate $Pr(A^c \bigcap B \bigcap S2|\text{a clone starts at t})$.

Assume L_t means the event that a clone starts at t. First we calculate

$$\begin{aligned} p1 &\triangleq Pr(S1|A^c, L_t) \\ &= \frac{\int_{-\infty}^{t} Pr(s + E_{sl} < t < t + E_{tl} < s + L_s - E_{sr} < s + L_s < t + L_t - E_{tr})ds}{\int_{-\infty}^{t} Pr(L_s \in A^c, L_t)ds} \\ &= \frac{\boldsymbol{E}\int_0^{\infty} G(u)[F(u + L_t - E_{tr}) - F(u + E_{sr} + E_{tl})]du}{\boldsymbol{E}\int_0^{\infty} G(u)[F(u + L_t - E_{tr}) - F(u + E_{sr} + E_{tl}) + 1 - F(u + L_t + E_{sr})]du} \end{aligned}$$

where $\boldsymbol{E}$, which is different from the E we defined in section 2.1, means get the operation to get the expectation of the corresponding expressions.

Using Yeh's (Yeh,1999) calculation of $Pr(A^c|L_t)$, we have

$$\begin{aligned} Pr(A^c \bigcap B \bigcap S1|L_t) \approx & p1 \times Pr(A^c|L_t) \times (2e^{-2\lambda E} - e^{-(2+\beta)\lambda E}), \\ Pr(A^c \bigcap B \bigcap S2|L_t) \approx & (1 - p1) \times Pr(A^c|L_t) \times (2e^{-2\lambda E} \\ & - e^{-4\lambda E + \int_{t+L_t-E_{tr}}^{t+L_t} [1 - G(s+L_s-E_{sr}-x)dxdY(t-s)]}), \end{aligned}$$

where $Y(\cdot)$ is the distribution of $t - s$, as was stated in Yeh's paper.

Therefore, replacing $Pr(A^c B|L_t)$ by the sum of the above two probabilities, we have our estimate of $p(t)$. Then we can get similar formula to estimate the number of scaffolds by replacing Yeh's $p(t)$ in equation (2) by ours. The curve labelled "ours" in Figure 1(a) uses our result.

2.3.2. *Average Scaffold Length.* In Port et al.'s paper (Port et al.,1995), they used greedy islands to describe scaffolds. They do not obtain a good estimation of the average scaffold number and scaffold length. Yeh used the following estimation:

$$E(\text{scaffold length}) \approx \frac{G(1-e^{-\lambda L})}{1-e^{-\lambda L}+N\{e^{-\lambda L}+e^{-7\lambda E}[e^{-\alpha\lambda E}-e^{-\lambda(L-3E)}](2-e^{-\beta\lambda E})\}}. \tag{3}$$

Notice the numerator in (3) is the part of genome that has been covered by clones and any region of the genome covered with two or more clones will be counted only once. So for two scaffolds interlacing but without overlapping at the ends, we should count their interlacing regions only once when we calculate the denominator of (3). That is why the denominator is similar to that in (2). Comparing (3) with (2), we can find that Yeh multiplied $Pr(A^cB|\text{a clone starts at t})$ by $e^{-2\lambda E}$ in (3). Yeh used the clones in A^c not to be extended. This idea is correct. Unfortunately, the argument omitted that some clones in A^c already can not be extended when calculating $Pr(B|A^c$ and a clone starts at t). Therefore, there is no need to multiply by $e^{-2\lambda E}$ for those clones. We believe it is better to use the following formula (4), (5) or (6). Figure 1(b) is the comparison of (3), (4), (5) and (6) with simulation results.

$$E(\text{length of scaffolds}) \approx \frac{G(1-e^{-\lambda L})}{1-e^{-\lambda L}+Ne^{-\lambda L}+Ne^{-5\lambda E}(e^{-\alpha\lambda E}-e^{-\lambda(L-3E)})} \tag{4}$$

$$E(\text{length of scaffolds}) \approx \frac{G(1-e^{-\lambda L})}{1-e^{-\lambda L}+N[e^{-\lambda L}+e^{-5\lambda E}(e^{-\alpha\lambda E}-e^{-\lambda(L-3E)})(1-e^{-\beta\lambda E})]} \tag{5}$$

$$E(\text{length of scaffolds}) \approx \frac{G(1-e^{-\lambda L})}{1-e^{-\lambda L}+Ne^{-\lambda L}+N(1-e^{-(3+\alpha)\lambda E})Pr(A^cB|L_t)} \tag{6}$$

(4)-(6) are all based on one principle: we don't want one scaffold contained completely in others to be counted. Note that our formula (4) is exactly what Yeh tried to get. As to formula (5), we count the scaffold beginning from L_t if the clone L_t can be extended. Our calculation strategy requires those scaffolds interlacing with the one beginning from L_t and starting before t can't be extended after L_t. Therefore, we should count the scaffold beginning from L_t since most clones of the scaffold are not contained by other scaffolds. If the scaffold beginning from L_t can not be extended, it contains only one clone and shares some parts of the clone with other scaffolds, and we have counted those in other scaffolds. So it is better for us to neglect the rest of the clone.

We multiply $Pr(A^cB|\text{a clone starts at t})$ by $1-e^{-(3+\alpha)\lambda E}$ in (6). Note that $1-e^{-(3+\alpha)\lambda E}$ is the probability that the clone L_t can be extended. If L_t can't be extended, it contains only one clone and some parts of the clone cover the same region as the clones in the previous scaffold. It is not a large error to neglect those scaffolds.

In summary, (4), (5) and (6) are better estimations than (3). Moreover, we can consider the event $S2$ as we did in previous section if we want to get even better result. But the argument is similar. So we will not write out here.

2.3.3. *The Average Number of Islands in a Scaffold.* Since each island can belong to only one scaffold, and we know how many scaffolds there are. In order to estimate how many islands there are in a scaffold on average, we have to estimate how many islands there are. Port's estimation (Port et al.,1995) is $2Ne^{-2\lambda E}$ while Yeh (Yeh,1995) claimed it was $1 + Ne^{-2\lambda E}$. It is certain that Port's answer is correct, but we should add $1 - e^{-\lambda E}$ as a correction due to the boundary effect. Later, we will show how to obtain $2Ne^{-2\lambda E}$ by using the distribution of gap size.

The following is the estimation of the average number of islands in a scaffold. (7) is Yeh's result while (8) is ours. Figure 3(a) gives a comparison between the two estimated.

$$E(\text{number of islands in a scaffold}) \approx \frac{1 + Ne^{-2\lambda E}}{(2)} \tag{7}$$

$$E(\text{number of islands in a scaffold}) \approx \frac{1 - e^{-2\lambda E} + 2Ne^{-2\lambda E}}{(2)} \tag{8}$$

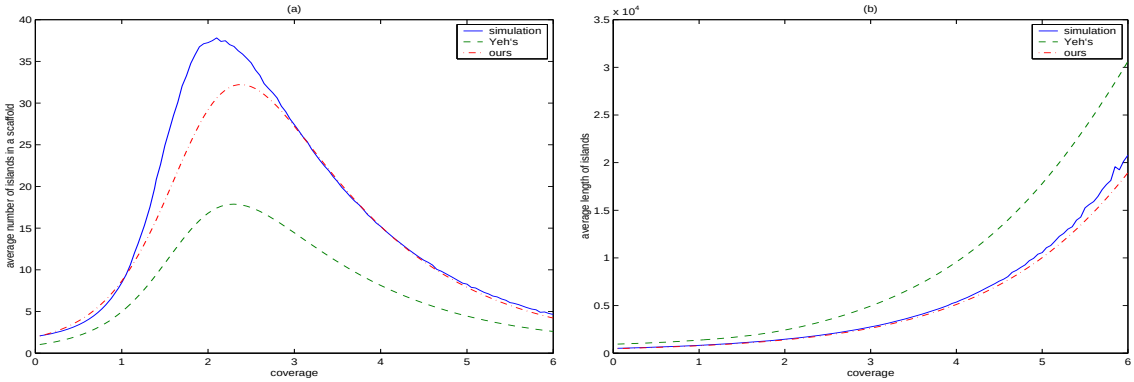

FIGURE 3. (a)Average number of islands in a scaffold as a function of coverage.(b)Average island length in a scaffold as a function of coverage.

2.3.4. *Average Island Length and Gap Size.* For any position in the genome, say t, the probability that no end covers t is $e^{-2\lambda E}$, where the factor of 2 is due to the fact that either left or right ends can cover t. Therefore, $G(1-e^{-2\lambda E})$ of the genome has been covered. (We can get also this formula from the distribution of gaps instead of Port et al.'s method). It is easy to obtain the following formula:

$$E(\text{length of islands}) \approx \frac{G(1 - e^{-2\lambda E})}{(2)(8)},$$

where the items in parentheses are the quantities in the corresponding equations. Based on the same argument, we have

$$E(\text{gap size in a scaffold}) \approx \frac{(5) - \frac{G(1-e^{-2\lambda E})}{(2)}}{(8) - 1}$$

For the gaps between adjacent islands, we obtain the following estimation.

$$E(\text{gap size between adjacent islands}) \approx \frac{Ge^{-2\lambda E}}{1 + (2N-1)e^{-2\lambda E}}$$

Figure 3(b) is a comparison of different estimations of the island length. Figure 4(a) is a comparison of different estimations of gap size between adjacent islands in a scaffold. Figure 4(b) is a comparison of different estimations of gap size between adjacent islands, where the curve labelled with "gapsize1" is the line for last formula above.

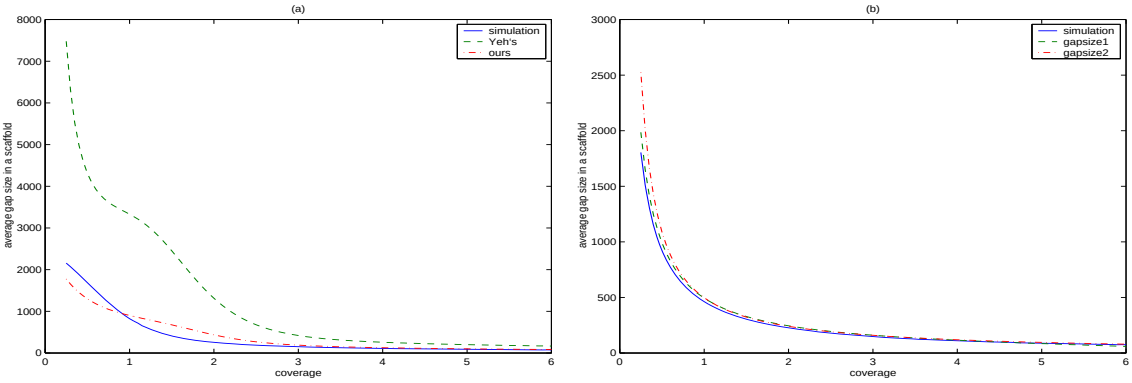

FIGURE 4. (a)Average gap size between adjacent islands in a scaffold as a function of coverage.(b)Average gap size between adjacent islands, which may belong to different scaffolds, as a function of coverage.

2.3.5. *Contig Scaffolds.* By contig scaffold, we mean those scaffolds composed of at least two clones. The counterpart of contig scaffold is singleton scaffold. By singleton scaffold, we mean those scaffolds containing only one clone. Define $r(t)$ as $Pr(\text{no other clone overlap with the ends of the clone } L_t | \text{a clone starts at t})$. It is easy to get $r(t) = e^{-(6+2\alpha)E}$, as can be found in Yeh, 1999. Therefore, the number of singleton scaffolds is $Nr(t) + (1 - e^{-\lambda L})r(t)$, where $(1 - e^{-\lambda L})r(t)$ is a correction for the boundary. And the length of genome that has been covered by singleton is $[Nr(t) + (1 - e^{-\lambda L})r(t)]L$. Therefore, we have the following formulae for contig scaffolds (Only formula (10) appears in Yeh (Yeh,1999)):

$$E(\text{number of contig scaffolds}) \approx (1 - e^{-\lambda L})[1 - r(t)] + N[p(t) - r(t)]. \tag{9}$$

$$E(\text{contig scaffold length}) \approx$$

$$\frac{G(1-e^{-\lambda L}) - [Nr(t) + (1-e^{-\lambda L})r(t)]L}{(1-e^{-\lambda L})[1-r(t)] + N[e^{-\lambda L} + e^{-5\lambda E}(e^{-\alpha\lambda E} - e^{-\lambda(L-3E)})(1-e^{-\beta\lambda E})]}. \tag{10}$$

(11)

$$E(\text{number of islands in contig scaffold}) \approx \frac{(1-e^{-2\lambda E})(1-r(t)) + 2Ne^{-2\lambda E}[1-r(t)]}{(9)}.$$

$$E(\text{length of islands in contig scaffold}) \approx \frac{G(1-e^{-2\lambda E}) - 2Nr(t)E}{(9)(11)}. \tag{12}$$

(13)

$$E(\text{gap size of contig scaffold}) \approx \frac{\frac{G(1-e^{-\lambda L})-[Nr(t)+(1-e^{-\lambda L})r(t)]L}{(9)} - \frac{G(1-e^{-2\lambda E})-2Nr(t)E}{(9)}}{(11)-1}.$$

The above formulae are good only when the coverage is larger than 3. Figure 5 contains the comparisons of those formulae with simulation results.

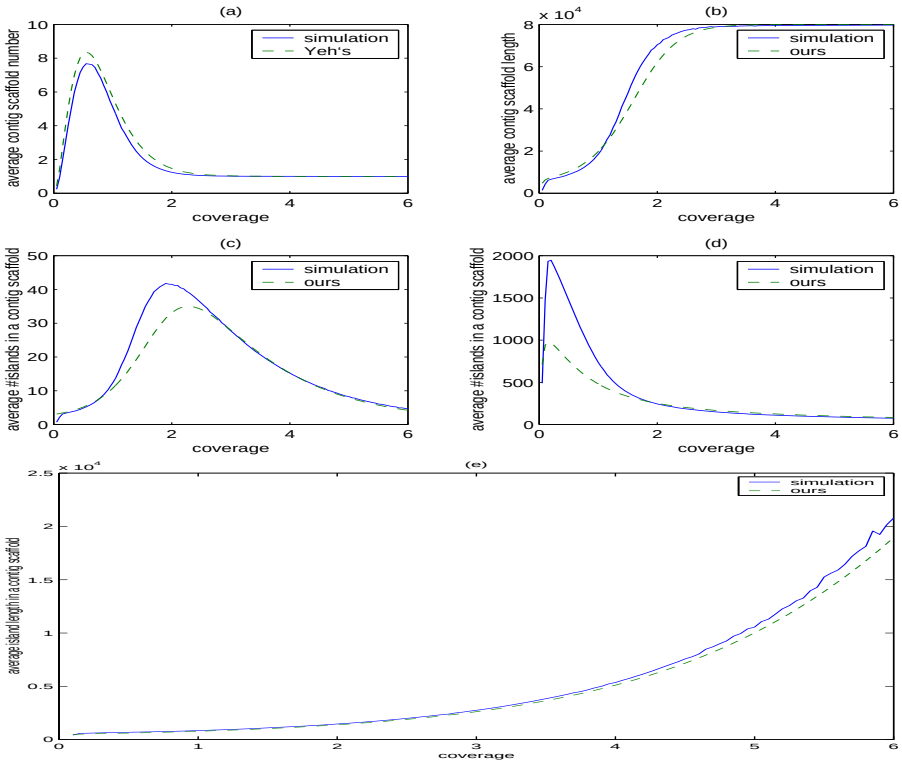

FIGURE 5. (a)Average number of contig scaffolds. (b)Average length of contig scaffolds. (c)Average number of islands in a contig scaffold. (d)Average gap size in a contig scaffold.(e)Average island length in contig scaffolds.

3. The Distribution of Sizes of Gaps between Adjacent Islands

There are two types of gaps. Gaps occur between adjacent islands or between adjacent islands in a scaffold. We are interested in the former here.

Assume $\bar{L}$ is the exact lower bound of the length of clones. Also assume the length of the ends is not larger than $\bar{L}$. Let A ={A gap beginning from 0 and with width $\geq x$ }, B={there is a clone ending at 0}. We want to calculate $Pr(A|B)$. For A occurring under the condition B, we need C1,C2 and C3, i.e.,

$$Pr(A|B) = Pr(\text{C1,C2 and C3}|B) = Pr(C1|B)Pr(C2|B)Pr(C3|B),$$

where

$C1 \triangleq$ {all clones with left ends starting in$(-\infty, -\bar{L})$can't have right ends overlapping with the interval$(0, x)$}

$C2 \triangleq$ {all clones with left ends starting in $(-\bar{L}, 0)$ can't have left ends overlapping with $(0, x)$ and must have right ends in $(x, +\infty)$}

$C3 \triangleq$ {no clone happens with left end beginning in $(0, x)$}

By using the thinning Poisson process idea, we calculate $Pr(C1|B)$ in the following way. We calculate the probability that the clone L_t has its right end overlap with $(0, x)$ for any $t < -\bar{L}$. We write this probability as $\acute{P}(t)$. Then

$$\begin{aligned}\acute{P}(t) =& Pr(t + L_t > 0, t + L_t - E_{tr} < x)\\ =& Pr(-t < L_t < x + E_{tr} - t)\\ \approx& \int [F(x + y - t) - F(-t)]dG(y)\end{aligned}$$

Therefore, we have $Pr(C1|B) \approx e^{-\lambda \int_0^\infty \int_{-\infty}^{-\bar{L}} [F(x+y-t)-F(-t)]dtdG(y)}$. Similarly, we have

$$\begin{aligned}Pr(C2|B) &\approx e^{-\lambda \int_0^\infty \int_{-\bar{L}}^0 \{1-G(-t)+G(-t)[F(x+y-t)-F(-t)]\}dtdG(y)};\\ Pr(C3|B) &\approx e^{-\lambda x};\end{aligned}$$

Recall what we calculated was the probability that there is a gap after a read, including the case of gap width 0. Therefore, in order to get the distribution of the gaps after an island, we have to multiply the above probabilities by a constant, which is the probability that there is no gap after the position zero given there is a clone ending at zero.

If we assume the length of the ends are fixed as E, we know $\bar{L}$ is $2E$. If we assume $x = 0$, we have

$$\begin{aligned}Pr(C1|B) \approx& e^{-\lambda \int_{\bar{L}}^{\bar{L}+E} [1 - F(t)]dt},\\ Pr(C2|B) \approx& e^{-2\lambda\bar{L} - \lambda \int_E^{2E} [1-F(t)]dt + \lambda \int_{\bar{L}}^{\bar{L}+E} [1-F(t)]dt},\\ Pr(C3|B) \approx& 1.\end{aligned}$$

Therefore,we know the probability there is no gap after a clone is $e^{-2\lambda E}$ in this case. Actually, this is always correct if the average length of ends is E.

We can use the above formula to calculate the average gap size. The line labelled with "gapsize2" in Figure 6 is the line.

4. Discussion and Conclusion

In this paper, we have made improvement in the formulae predicting aspects of double-end (or mate-pair) clone DNA sequence assembly. In particular, we present formulae for the average number and length of scaffolds, islands, and gaps; and estimate the average number of islands in a scaffold, and so on. In addition, we estimate the distribution of the gap size between adjacent islands.

It is necessary to make approximations. We and Yeh both assume there is only one clone in A^c when calculating the average number of scaffolds. From simulations, we know the probability that two scaffolds share a region with width longer than the length of one clone is very small, when coverage is larger than 2.5. In future work we hope to add rigorous bounds on this and other approximation errors.

In our calculation of average scaffold number, we included clones such as L_{S2} instead of only the clones such as L_{S1}; see Figure 2(a) and Figure 2(c). From Figure 1(a), we can only see a little improvement. But if we use a mixture of very different clone sizes, such as Celera used, in our preliminary simulations, our result seem to be much better than Yeh's. We will study this further.

In summary, our estimators appear to behave well. We have not checked our model against the output of sequence assemblers because assembled sequence is the result of numerous and complex operations. This is in our future plans.

It was our goal to provide guidance for the design of sequencing project, and we have not yet fully succeeded in this. Because assembly depends on more complex details of the island-scaffold structure than what we have obtained, our future success will depend on further analysis.

5. Acknowledge

Many thanks to Granger Sutton and Gene Myers for suggesting this work. X. Li is grateful to the university of Southern California for University Fellowship support, and both X. Li and MSW were partially supported by NIH grant 53-4855.

References

1. Port E., Sun F., Martin D., and Waterman M.S.(1995) Genomic mapping by end-characterized random clones: a mathematical analysis. Genomics 26,84-100.
2. Roach J., Boysen D., Wang K., and Hood L.(1995) Pairwise end sequencing: A unified approach genomic mapping sequencing. Genomics 26,345-353.
3. Siegel A.F., Van Den Engh G., Hood L., Trask B., and Roach J.(2000) Modeling the feasibility of whole genome shotgun sequencing using a pairwise end strategy.Genomics 68,237-246.
4. Venter J.C. et al. (2001) The sequence of human genome. Science,vol291,16 Feb, 1304-1351.
5. Venter J.C. et al. (1998)Shotgun sequencing of the human genome. Science, vol280, 1540-1542.

6. Waterman M.S. (1995) Introduction to Computational Biology: Maps, sequences and genomes. Chapman and Hall.
7. Webber J., Myers E.(1997) Genome Research. May;7(5):401-409.
8. Yeh R. (1999) Statistical issues in genomic mapping and sequencing. Dissertation. UC Berkeley.