JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 47–57

The Effect of Nonhomogeneous Clone Length Distribution on the Progress of an STS Mapping Project

SOPHIE SCHBATH,[1] NATHALIE BOSSARD,[1] and SIMON TAVARÉ [2]

Abstract

We provide both theoretical and simulation results on the progress of an STS mapping project in the presence of clone length inhomogeneity. For an example in which the genome comprises alternating regions of clones with short and long average length, the main conclusion is that the efficiency of the project is clearly decreased in the presence of such inhomogeneity. The case of deterministic clone length gives the worst progress. The general simulation algorithm we propose shows that strategies that space the anchors as regularly as possible do best: fewer contigs of larger average length are expected. The simulation algorithm can be used to study many statistical properties of the progress of any anchoring project.

Key words: STS mapping, nonhomogeneity, anchored islands, genome coverage.

1. INTRODUCTION

An STS mapping project of a genome consists of constructing a set of ordered and overlapping genomic fragments, called clones, spanning the entire genome. This is done by assembling into contigs (or anchored islands) clones that contain an STS, say an anchor, in common. Clones and anchors are chosen from libraries, and produce contigs which cover some regions of the genome. The progress of such an STS mapping project is often described by the number of anchored islands, the typical length of an anchored island and the proportion of genome not covered by anchored islands.

Simple models of anchoring assume that we have a perfectly representative library of genomic clones. In practice, physical mapping data reveal that clones do not occur homogeneously along the genome and clone lengths have distributions that vary with position in the genome. These effects are due to cloning bias. Since fragments are produced by partial digestion of the genome, the clone locations are correlated with restriction sites that are known to be nonhomogeneously distributed along the genome. The nonhomogeneity of clone locations decreases the efficiency of the mapping project (Schbath, 1997): the mean proportion of genome covered by anchored islands is smaller than in an homogeneous design, there are more anchored islands and they are smaller. The effect of having a genome composed of two kinds of homogeneous regions, those

[1]Institut National de la Recherche Agronomique, Unité de Biométrie, 78352 Jouy-en-Josas, France. E-mail: Sophie.Schbath@jouy.inra.fr

[2]Program in Molecular Biology, Department of Biological Sciences, SHS172, University of Southern California, Los Angeles, California 90089-1340. E-mail: stavare@hto.usc.edu

with few clones and those with many clones (the so-called hotspot model) has been extensively studied in Ewens *et al.* (1996) and Schbath (1996). Moreover, it has been noticed in YAC maps of human chromosomes that YACs tend to be shorter in regions of relatively high G+C content than in regions of low G+C content. The reason for this is thought to be that *S. cerevisiae*, the host in which YACs are grown, has a low G+C content (35%). Presumably *S. cerevisiae* has a nucleotide metabolism adapted to replicating low G+C DNA strands, and so does not readily tolerate cloned DNA with a high G+C content. It is then natural to think that regions of short clones have a low depth of coverage. This lack of efficiency could be balanced by a clone hotspot: one needs more clones in regions where clones tend to be shorter.

In this paper we study the progress of an STS mapping project in the presence of clone length inhomogeneity. We still assume random anchoring, the anchors being homogeneously distributed across the genome. In equations (1)–(3) we give formulae, based on earlier work of Schbath (1997), for the mean number of anchored islands and the mean proportion of the genome covered by oceans. A formula for the mean length of an anchored island is given in a special case. The results are illustrated for representative parameter values. The main conclusions are that the efficiency of the mapping project is clearly decreased in presence of clone length inhomogeneity and that using clones with same fixed length does not provide a better strategy. There are many interesting statistical properties of the progress of a mapping project that have yet to be derived theoretically. To study some of these, we provide simulations based on a dynamic algorithm described in more detail in the appendix. This algorithm, which takes as input any stream of clone and anchor positions, is particularly useful for analyzing clone and anchor distributions for which no theoretical results are yet available. For instance, we show that regularly spaced STSs seem to be a good strategy. The program is available at `http://www-bia.inra.fr/J/AB/genome/ISLAND/welcome.html`.

2. PROPERTIES OF ANCHORED ISLANDS

2.1. Analytical results

In this section we assume that clones are distributed independently across the genome according to a Poisson process, their lengths having a distribution that depends on the location of the clone. For definiteness, we use the right-hand end of a clone to define its location. The anchors are also supposed to be randomly distributed along the genome according to a Poisson process. We use the following notation:

G	genome length in basepairs (bp),
L_i	length of clone ending at position i on the genome, in basepairs; it may be constant or variable,
$L_{\max}$	maximal mean length of clones, in basepairs,
N	(mean) number of clones in the library,
M	(mean) number of anchors studied.

For convenience, we rescale the genome by $L_{\max}$ so that the normalized genome is represented by the segment $(0, g]$ with $g = G/L_{\max}$. Thus, $\alpha := N/g$ and $\lambda := M/g$ represent, respectively, the mean number of clones and the mean number of anchors per unit in $(0, g]$ and are precisely the rates of the associated Poisson processes. Q_t denotes the normalized length of a clone ending at t, $t \in (0, g]$. We consider first the general framework where the normalized clone length Q_t is distributed according to a probability density function f_t.

We focus on three quantities of interest: the mean number of anchored islands, the mean length of an anchored island and the mean proportion of oceans (that is, genome not covered by anchored islands). First of all, we define the auxiliary functions

$$\mathcal{F}_t(w) = \mathbb{P}(Q_t \geq w) = \int_w^\infty f_t(q)dq$$

and

$$J(t; x) = \exp\left(-\alpha \int_x^\infty \mathcal{F}_{t+u}(u)du\right),$$

which depend on α and f_t. The function $J(t; x)$ represents the probability that no clone covers simultaneously the points t and $t+x$ $(x > 0)$. The result below may be proved using the methods developed in Schbath (1997), and we omit the details here. For the interested reader, they are available at `http://www-bia.inra.fr/J/AB/genome/ISLAND/welcome.html`.

Proposition 1 *With the notation above,*
(i) the mean number of anchored islands ending in $(0, g]$ *is*

$$\alpha\lambda \int_0^g \int_0^\infty \mathcal{F}_t(w) J(t-w; w) e^{-\lambda w} dw\, dt, \tag{1}$$

(ii) the mean proportion of oceans is

$$\frac{\lambda^2}{g} \int_0^g \int_0^\infty \int_0^\infty \frac{J(t; v) J(t-w; w)}{J(t-w; v+w)} e^{-\lambda(v+w)} dw\, dv\, dt, \tag{2}$$

(iii) if there exists θ *such that* $f_t \equiv f_{t+\theta}$ *for all* $t \in \mathbb{R}$, *then the mean length of an anchored island is*

$$\frac{\int_0^g \Omega(t)dt}{\textit{mean number of anchored islands}} \times L_{max} \tag{3}$$

where

$$\Omega(t) = 1 + \lambda^2 \int_0^\infty \int_0^\infty [J(t-w; v+w) - J(t; v) - J(t-w; w)] e^{-\lambda(v+w)} dw\, dv\,.$$

2.2. Application

In this section we apply the results in (1)–(3) to a case in which long and short clones alternate, reflecting, for instance, high/low G+C content bands in the genome. There are only two clone length distributions, one for the short clones and one for the long clones. We assume long clones have a normalized length uniformly distributed in $[1-s;1+s]$, $s \in [0,1]$, whereas small clones have a normalized length uniformly distributed in $[L_{\min}/L_{\max} - s; L_{\min}/L_{\max} + s]$, where $L_{\min}$ denotes the minimal mean length in basepairs; here this is the mean length of small clones. Since a normalized clone is no longer than 2, note that for $w \geq 2$, $\mathcal{F}_t(w) = 0$ and $J(t;w) = 1$; it implies in particular that all the integrals defined above are in fact over finite intervals, facilitating their numerical calculation.

As the starting point in our numerical investigations, the parameter values corresponding to the first mapping project of *A. thaliana* have been used, as in Ewens *et al.* (1991): a genome length of 100Mb, 2300 clones of 250kb and 500 anchors. In our case, $L_{\min}$ and $L_{\max}$ will not be necessarily equal to 250kb but we always preserve the relation $L_{\min} + L_{\max} = 500$kb. We chose arbitrarily to split the genome into 20 regions of 5 Mb: 10 regions of long clones alternating with 10 regions of short clones. The results shown below are not affected by the number of regions.

For a fixed value of s, Figure 1 clearly shows that inhomogeneity in the clone length reduces the efficiency of the mapping project. Indeed, by increasing the mean length of long clones ($L_{\max}$) we obtain more anchored islands, but they become so short that the mean proportion of oceans increases. The trends are the same for different values of s, but it is worthwhile noting that having clones with deterministic length ($s = 0$) is not a good strategy. This has already been pointed out in the homogeneous framework, but it is not a universal rule for every clone length distribution (see Arratia *et al.* (1991)).

In the remainder of the paper, we take $s = 100\text{kb}/L_{\max}$, so that short clones have a length uniformly distributed in $[L_{\min} - 100\text{kb}, L_{\min} + 100\text{kb}]$, whereas the length of long clones is uniformly distributed in $[L_{\max} - 100\text{kb}, L_{\max} + 100\text{kb}]$.

It is obvious that the more clones and the more anchors one studies, the better is the physical map. However, the efficiency has to be balanced with the cost of the project. When increasing the mean numbers of clones and anchors, we obtain curves that are similar to those in Arratia *et al.* (1991). For instance, if short (respectively, long) clone lengths are uniformly distributed in $[50, 250]$kb (respectively, $[250, 400]$kb) — corresponding to $L_{\max} = 350$kb and $L_{\min} = 150$kb — Figure 2 shows that it is not necessary, in terms of genome covered, to consider more than 1100 anchors and 1500 randomly generated clones (we obtain 90.3% coverage in 185 contigs of 526kb mean size, as shown in Table 1). From these limits, the gain from the random approach is negligible and directed approaches should be used, such as considering clone ends or contig ends as STS. This fact has been taken into account in recent mapping projects (cf. Nelson and Speed (1994); Port *et al.* (1995); Schmidt *et al.* (1996); Nagaraja *et al.* (1997); and Bouffard *et al.* (1997)).

2.3. Simulations

As in most theoretical analyses of mapping projects, only mean values of quantities of interest have been found explicitly. It is also of interest to have the variance of these quantities; simulation seems to be the only viable approach at present. In this section we give some simulation results that are applications of our algorithm that calculates the number of anchored islands, the average length of the anchored islands and the proportion of oceans, with respect to the positions of the

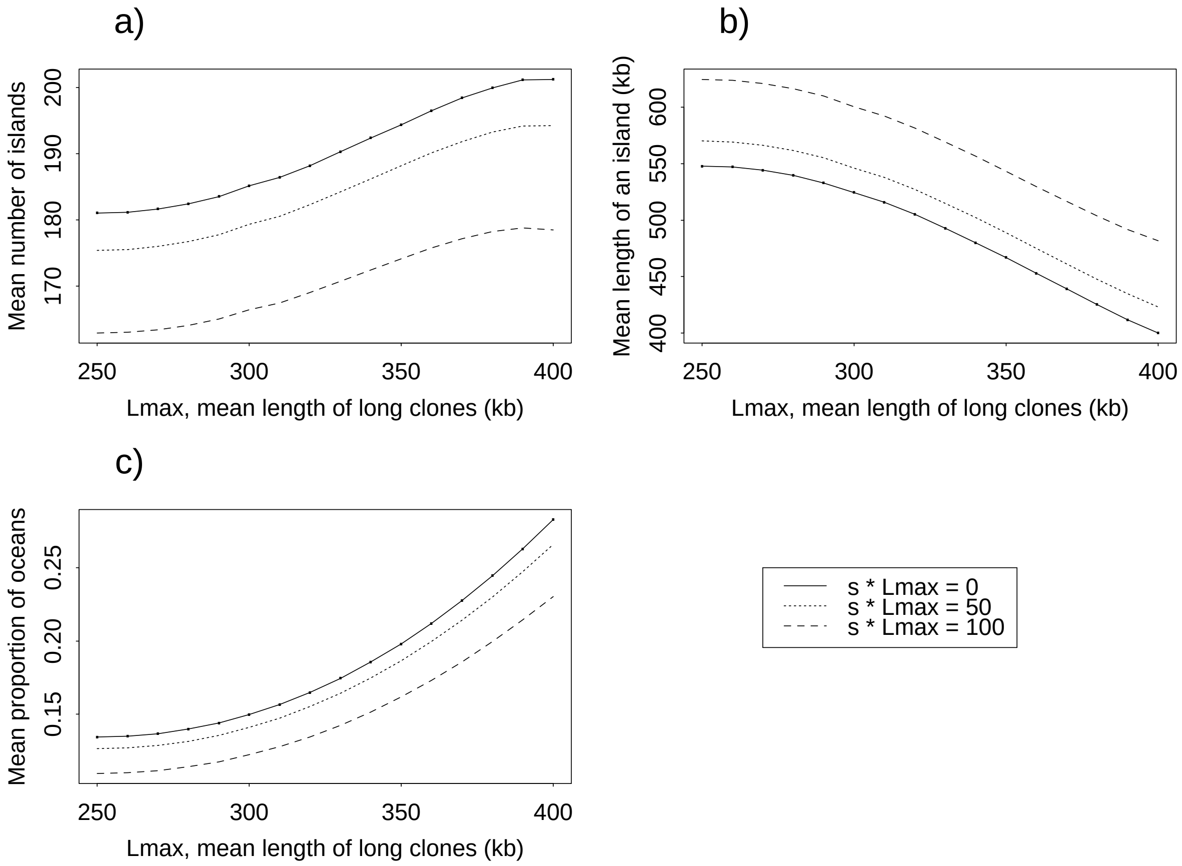

Figure 1: Variation of the mean number of anchored islands (**a**), the mean length of an anchored island (**b**) and the mean proportion of oceans (**c**) when varying the mean length of long clones ($L_{\max}$) from 250kb to 400kb. The results have been obtained with a genome of length 100Mb split into 20 regions of small/long clones, 2300 clones and 500 anchors. The length of long clones is uniformly distributed in $[L_{\max} - sL_{\max}, L_{\max} + sL_{\max}]$ whereas small clones have a length uniformly distributed in $[500 - L_{\max} - sL_{\max}, 500 - L_{\max} + sL_{\max}]$. Each of the three quantities (**a**, **b** and **c**) have been calculated for three different values of $sL_{\max}$: $sL_{\max} = 0$ (solid line), $sL_{\max} = 50$kb (dotted line) and $sL_{\max} = 100$kb (dashed line).

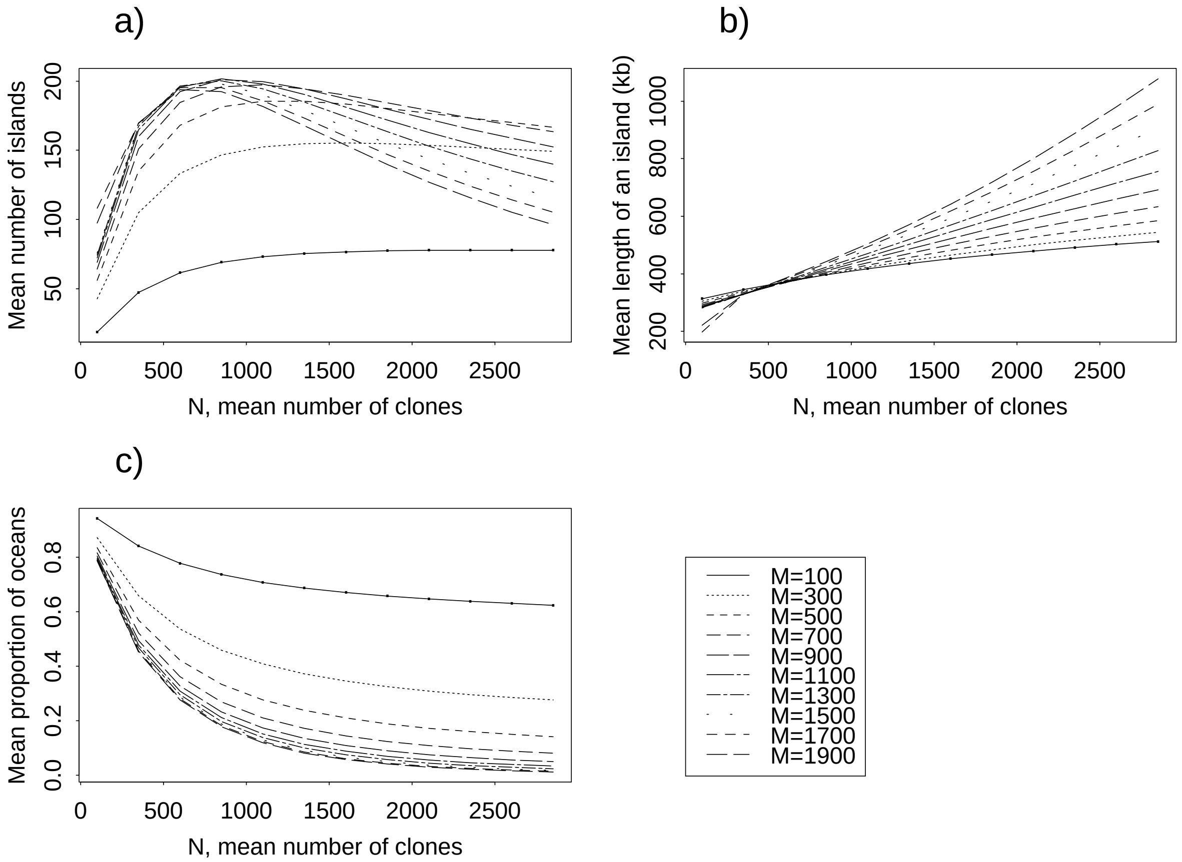

Figure 2: Variation of the mean number of anchored islands (**a**), the mean length of an anchored island (**b**) and the mean proportion of oceans (**c**) when varying simultaneously the mean number of clones from 100 to 2800 (x-axis coordinate) and the mean number of anchors from 100 to 1900 (different style line curves). The results have been obtained with a genome of length 100Mb split into 20 regions of small/long clones. Small clones have a length uniformly distributed in $[50, 250]$kb whereas the length of long clones is uniformly distributed in $[250, 450]$kb.

clones and the anchors along the genome.

The basic dynamic algorithm is described in detail in the appendix. It can be extended easily to study various other quantities of interest not yet available through a mathematical analysis. Moreover, the algorithm has been designed in such a way that one can simply specify the clone and anchor locations along the genome. Clones (respectively, anchors) are considered one after another starting from the right-hand end of the genome to the left-hand end. The positions of the anchors and of both ends of the clones can either be simulated according to a specified model or read from previously generated input files. Both versions are available in the `ISLAND` program available at `http://www-bia.inra.fr/J/AB/genome/ISLAND/welcome.html`; only the model used in Section 2.2 is implemented in the first version.

In the following simulations, we consider exactly the same model as in the previous section, so that we can compare the theoretical results with the simulated ones. We therefore used homogeneous Poisson processes for the anchor locations and the right-hand ends of clones, and two uniform distributions for the clone length depending on the left-hand end position of the clone along the genome (which is split into 20 regions of 5Mb: 10 regions where the average clone length is $L_{\max}$ alternated with 10 regions where the average clone length is $L_{\min}$).

We have compared, for different values of $L_{\max}$ and N, the theoretical mean of the three quantities of interest obtained from equations (1)–(3) with the associated empirical average calculated over 100 iterations of the algorithm (the so-called simulated mean). Figure 3 shows that the simulated results are very close to the theoretical ones. Moreover, as we noted earlier, simulations allow us to obtain estimates of variances and then confidence intervals. Table 1 gives, for instance, the theoretical mean of the three quantities of interest and the associated 95% confidence interval calculated over 100 simulations when the length of small (respectively, long) clones is uniformly distributed in $[50, 250]$kb (respectively, $[250, 450]$kb); two cases have been considered: 1500 clones and 1100 anchors, and 2300 clones and 500 anchors.

Table 1: COMPARISON OF THE 95% CONFIDENCE INTERVAL CALCULATED OVER 100 SIMULATIONS FOR SEVERAL QUANTITIES OF INTEREST WITH THE THEORETICAL MEAN GIVEN BY EQUATIONS (1)–(3)

	1500 clones 1100 anchors		*2300 clones 500 anchors*	
	Theoretical mean	*95% confidence interval*	*Theoretical mean*	*95% confidence interval*
Number of anchored islands	185.53	184.79± 1.63	174.69	175.02 ± 1.57
Length of an anchored island (kb)	526.49	528.30± 4.19	541.35	540.57± 3.78
Proportion of oceans	0.0966	0.0984± 0.0018	0.1620	0.1633± 0.0035

These results were obtained with a genome of length 100Mb split into 20 alternating regions of long/short clones. Short clones have a length uniformly distributed in $[50, 250]$kb whereas the length of long clones is uniformly distributed in $[250, 450]$kb.

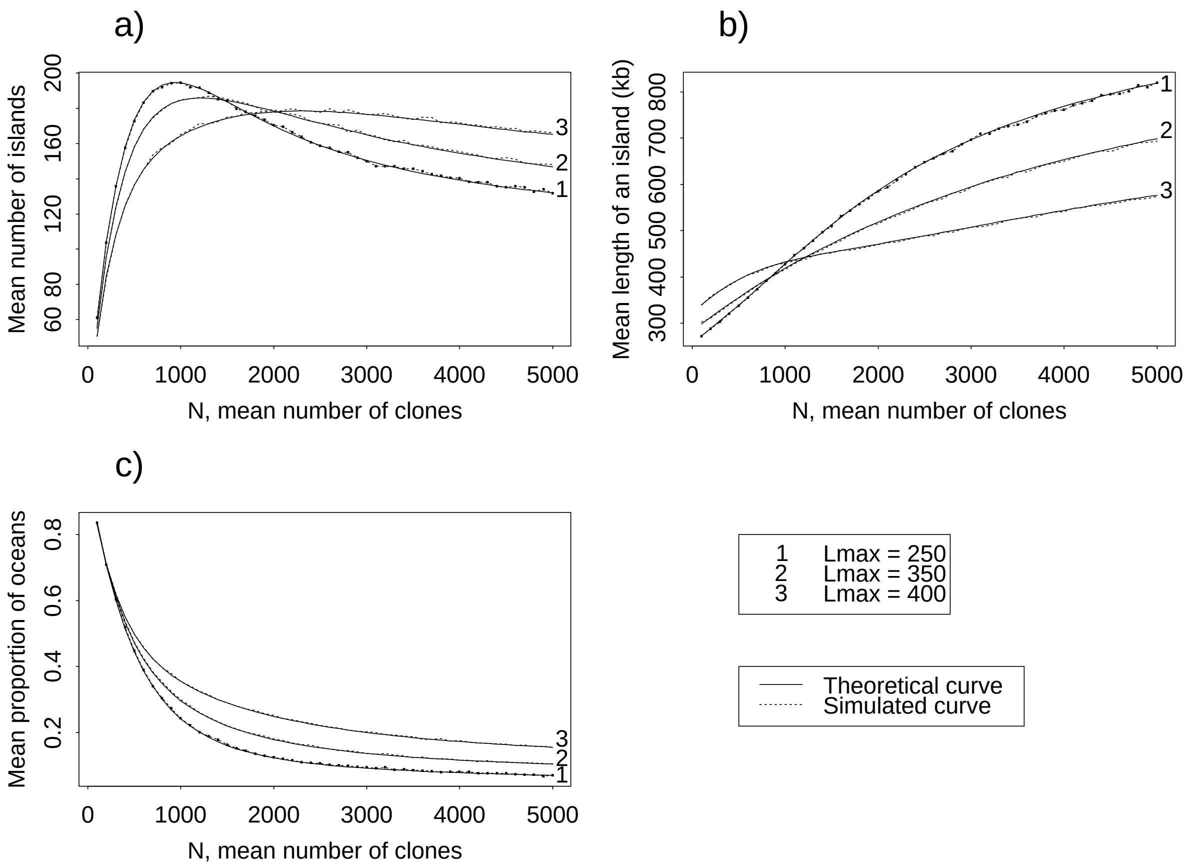

Figure 3: Comparison of the theoretical mean and the simulated mean of the number of anchored islands (**a**), the length of an anchored island (**b**) and the proportion of oceans (**c**). The results have been obtained with a genome of length 100Mb split into 20 alternated regions of small/long clones, 500 anchors and 100 to 5000 clones. Long clones have a length uniformly distributed in $[L_{\max} - 100, L_{\max} + 100]$ whereas the length of small clones is uniformly distributed in $[400 - L_{\max}, 600 - L_{\max}]$. We have considered the three following cases: $L_{\max} = 250$kb (1), $L_{\max} = 350$kb (2) and $L_{\max} = 400$kb (3). Solid lines correspond to theoretical means calculated from equations (1)–(3), whereas dashed lines correspond to simulated means.

3. DISCUSSION

It is now well known that GC-rich regions of chromosomes are difficult to clone in YACs (Bernardi (1995); Saccone *et al.* (1996)), leading to short YACs in these parts of the chromosome. The poor YAC coverage of GC-rich regions observed by Bouffard *et al.* (1997) can then be clearly explained by our result since we showed that inhomogeneity in the clone length distribution along the genome decreases the efficiency of the physical mapping project. The identification of the GC-richest bands in the chromosome to be mapped would then allow us, using the results in (1)–(3) or the simulation algorithm described in this paper, to predict the progress of a YAC-based STS-content mapping project more accurately than by using results derived under homogeneity assumptions for the clone lengths.

In the special case in which the genome is divided into alternating regions of short and long clones, as described in Section 2.2, a rough estimate of the progress of the mapping project can be obtained by applying results from Arratia *et al.* (1991) separately on each of the alternating regions. Any error in this approach will be due to boundary effects. To assess the likely magnitude of such effects, we considered the case given in the second scenario in Table 1: a genome of 100Mb divided into regions of alternating long and short clones, using 2300 clones and 500 anchors. We considered the cases of 20 regions and 100 regions, with both uniformly distributed clone lengths (as in Table 1) and deterministic clone lengths with the same means. The results are presented in Table 2. We note that the heuristic results do not depend on the number of regions, in contrast to the theoretical results. We also see from Table 2 that the heuristic works well when there are few alternating regions, and less well as the number of regions increases.

Table 2: Comparison of Heuristic and Theoretical Results (1)–(3) for Alternating Regions of Short and Long Clones

	Uniform clones		*Deterministic clones*	
	Theoretical mean	*Heuristic mean*	*Theoretical mean*	*Heuristic mean*
20 regions				
Number of anchored islands	174.69	175.94	194.97	196.15
Length of an anchored island (kb)	541.35	535.09	465.19	460.46
Proportion of oceans	0.1620	0.1665	0.1983	0.2033
100 regions				
Number of anchored islands	169.56	175.94	189.90	196.15
Length of an anchored island (kb)	566.62	535.09	485.28	460.46
Proportion of oceans	0.1437	0.1665	0.1787	0.2033

The results were obtained with a genome of length 100Mb, 2300 clones, 500 anchors and two clone length scenarios: short clones uniformly distributed in $[50, 250]$kb and long clones uniformly distributed in $[250, 450]$kb, and the deterministic case with the same means as above.

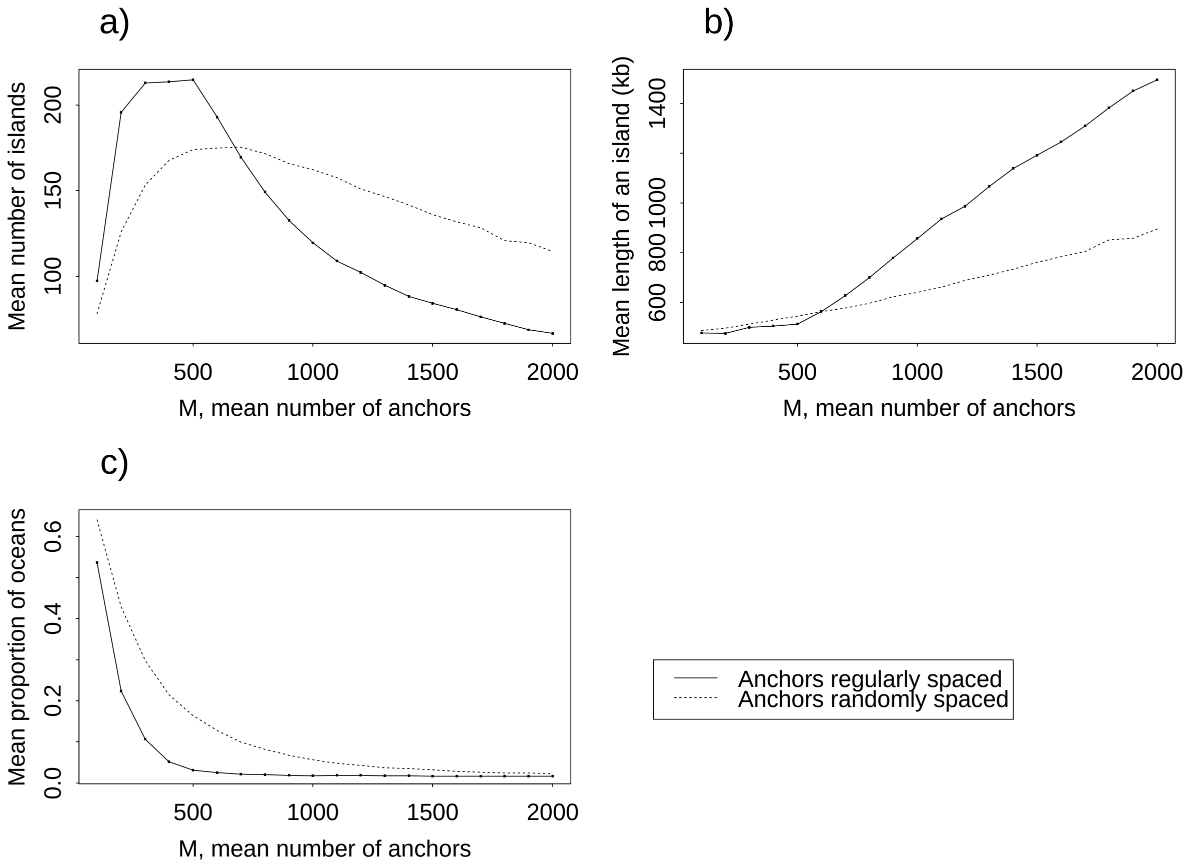

Figure 4: Comparison of the mean numbers of anchored islands (**a**), of the mean lengths of an anchored island (**b**) and of the mean proportion of oceans (**c**) in the two following cases: the anchors are randomly distributed along the genome (dashed lines) and the anchors are regularly spaced along the genome. The results have been obtained with a genome of length 100Mb split into 20 alternated regions of small/long clones, 2300 clones and 100 to 2000 anchors. Long clones have a length uniformly distributed in $[250, 450]$kb whereas the length of small clones is uniformly distributed in $[50, 250]$kb.

Many recent mapping projects tend to provide a map with high STS resolution (cf. Hudson *et al.* (1996); Bouffard *et al.* (1997); Nagaraja *et al.* (1997)). We used our simulation algorithm to compare the quality of the maps when using anchors regularly spaced along the genome and using a homogeneous Poisson process that can be thought of as anchors uniformly located along the genome (see Figure 4). As might be anticipated, our conclusion is that trying to obtain regularly spaced STSs seems to be a better strategy. From Figure 4, we can see that the mean proportion of oceans is smaller when the anchors are regularly spaced. What seems much less obvious is that when increasing the number of anchors, the mean proportions of oceans tend to be similar in both cases but using regularly spaced anchors gives far fewer contigs that are much longer on average. This may be preferable in practice.

APPENDIX

We describe the algorithm used to calculate the number of anchored islands, the average length of the anchored islands and the proportion of oceans, with respect to the positions of the clones and of the anchors along the genome. This algorithm is dynamic in that the quantities of interest are dynamically updated without storing the locations of all the clones and all the anchors. Only the positions of the current clone and the current anchor are needed at a given step. The functions `Generate-an-anchor` and `Generate-a-clone` provide the positions of the next clone and the next anchor, going from the right-hand end to the left-hand end of the genome. These positions can either be simulated as required or read in from previously generated input files. In what follows, the beginning of a clone (or an anchored island) refers to its left-hand end; the end of a clone (or anchored island) to its right-hand end.

If the positions of the clones and of the anchors are simulated, the genome length (`G`) and the maximal mean length of clones (`Lmax`) are input as parameters to the algorithm; otherwise, in addition to the positions of the clones and of the anchors, only `G` is input as a parameter to the algorithm (and `Lmax` is then set to 1). Before giving the main lines of the algorithm (Figure 5), we list the variables used in it:

Position marks on the genome:

- `Anc`: position of the current anchor,
- `BegClo`: position of the beginning of the current clone,
- `EndClo`: position of the end of the current clone,
- `BegIsl`: position of the beginning of the current anchored island,
- `EndIsl`: position of the end of the current anchored island,
- `BegFolIsl`: position of the beginning of the anchored island just following (to the right) the current anchored island;

Quantities of interest:

- `NbIsl`: current number of anchored islands,
- `LgIsl`: average length of the anchored islands,
- `Ocean`: proportion of oceans;

Intermediate quantities of interest:

- `TotLgIsl`: current total length of the anchored islands,
- `TotOcean`: current total length of the oceans;

Boolean variable:

- `IslFlag`: flag to know whether the current island has been completed and a new one may start (`IslFlag`= 0) or the current island may still be extended (`IslFlag`= 1).

At the beginning of the algorithm, the positions of the current clone, of the current anchor, of the current anchored island and `BegFolIsl` are set to `g` (the right-hand end of the normalized genome), and the flag `IslFlag` is set to 0 with the current quantities of interest. However, because the current quantities of interest are only updated when the current anchored island has just been completed (meaning that one has found the real beginning of the current island), the current number of anchored islands is initialized to -1. Then, one generates the position of the current clone (`BegClo` and `EndClo`) and positions of anchors (`Anc`) until the current anchor occurs before the end of the clone ($\texttt{Anc} \leq \texttt{EndClo}$). There are then two possibilities: either the current anchor is located before the beginning of the current clone ($\texttt{Anc} < \texttt{BegClo}$) or the current clone is anchored by the current anchor ($\texttt{BegClo} \leq \texttt{Anc} \leq \texttt{EndClo}$). In the first case, one generates new clones until the current clone begins before the current anchor ($\texttt{BegClo} \leq \texttt{Anc}$). If the current anchor is now located after the end of the current clone ($\texttt{Anc} > \texttt{EndClo}$), then all the previous steps are done again, so that we are now assured that the current clone is anchored by the current anchor. The first time this event occurs, it means that we start the first anchored island; `IslFlag` is still equal to 0 so the number of island, `NbIsl`, becomes 0, the current island is set to the current clone (`BegIsl:=BegClo` and `EndIsl:=EndClo`), and `IslFlag` is set to 1 since the current island may be extended from the left. We then generate the position of a new clone and start the main loop again. Only two events can occur: either the current island will be completed (because the next anchored clone will not be anchored to the current island), or the current island will be extended from the left because the next anchored clone will be anchored to the current island. We describe the following two cases since they cover the entire strategy in the algorithm. Assume one has generated anchors and clones until the current clone is anchored by the current anchor ($\texttt{BegClo} \leq \texttt{Anc} \leq \texttt{EndClo}$), as we did at the very beginning of the algorithm.

- If the current anchored clone is anchored to the current island ($\texttt{Anc} \geq \texttt{BegIsl}$), `IslFlag` is still equal to 1 and the current island is extended to the left until the beginning of the current clone (`BegIsl:=BegClo`). A new clone is generated.

- If the current anchored clone is not anchored to the current island ($\texttt{Anc} < \texttt{BegIsl}$) then the current island is now complete and a new one will start: `IslFlag` is set to 0. Since the current island is complete, one updates the quantities of interest:

  - the number of anchored islands is increased by 1,
  - the total length of anchored islands is increased by the length of the current anchored island, that is $\texttt{EndIsl} - \texttt{BegIsl}$,
  - and, if the current island does not overlap the following (to the right) anchored island (if $\texttt{EndIsl} < \texttt{BegFolIsl}$), then the total length of oceans is increased by the distance between these two anchored islands.

  The position marks of the current anchored island and of the anchored island that just follows it from the right are shifted to the left: `BegFolIsl` is set to `BegIsl` and the current island is set to the current clone. Finally, `IslFlag` is set to 1 (since the current island may be extended) and a new clone is generated.

This process is continued until the current clone or the current anchor falls out of the genome ($\texttt{Anc} < 0$ or $\texttt{BegClo} < 0$). As soon as one of the stop conditions occurs, we need to update for

```
g := G/Lmax
Anc := g;      BegIsl := g;     NbIsl:= -1;   TotLgIsl:= 0
BegClo := g;   EndIsl := g;     LgIsl := 0;   TotOcean:= 0
EndClo := g;   BegFolIsl := g;  Ocean := 0;   IslFlag := 0
Generate-a-clone
Generate-an-anchor
While ((Anc >= 0) and (BegClo >= 0)) do
      While (Anc > EndClo) do
            Generate-an-anchor
            If (Anc < 0) then break
      EndWhile
      If (Anc > 0) then
         If (Anc < BegIsl) then
            IslFlag:=0     // current island is now complete
         EndIf
         While (Anc < BegClo) do
            Generate-a-clone
            If (BegClo < 0) then break
         EndWhile
         If ((BegClo > 0) and (Anc < EndClo)) then
            // the current clone is anchored, then two cases:
            If (IslFlag = 0 ) then
               // if the current island is complete, one needs to
               // update the current quantities of interest:
               NbIsl := NbIsl + 1
               TotLgIsl := TotLgIsl + EndIsl - BegIsl
               If (BegFolIsl > EndIsl) then
                  TotOcean := TotOcean + BegFolIsl - EndIsl
               EndIf
               BegFolIsl := BegIsl
               BegIsl := BegClo    // the current clone becomes
               EndIsl := EndClo    // the new current island
               IslFlag := 1        // current island may now be extended
            Else
               // if the current island is not yet complete,
               // it is extended from the right
               BegIsl := BegClo
            EndIfElse
            Generate-a-clone
         EndIf
      EndIf
EndWhile
// Updating of the current quantities of interest since
// the current (last) island has not yet been taken into account
If (BegFolIsl > EndIsl) then
   TotOcean := TotOcean + BegFolIsl -EndIsl
EndIf
TotLgIsl := TotLgIsl + EndIsl - BegIsl
NbIsl := NbIsl + 1
TotOcean := TotOcean + BegIsl
LgIsl := (TotLgIsl/NbIsl)*Lmax
Ocean:= TotOcean/g
```

Figure 5: Algorithm to calculate properties of a genomic map with anchored clones

the last time the quantities of interest because the current (and last) anchored island has not yet been taken into account. The proportion of oceans and the average length of the anchored islands are then computed from the total length of oceans and the total length of anchored islands.

ACKNOWLEDGEMENTS

We thank Eric Green, Phil Green, Eric Lander, and David Schlessinger for helpful comments about the inhomogeneity of clone lengths. This work was supported in part by NSF grant DBI 95-04393.

REFERENCES

Arratia, R., Lander, E. S., Tavaré, S. and Waterman, M. S. (1991). Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* **11:** 806–827.

Bernardi, G. (1995). The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* **29:** 445–476.

Bouffard, G., Idol, J., Braden, V., Iyer, L., Cunningham, A., Weintraub, L., Touchman, J., Mohr-Tidwell, R., Peluso, D., Fulton, R., Ueltzen, M., Weissenbach, J., Magness, C. and Green, E. (1997). A physical map of Human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Research* **7:** 673–692.

Ewens, W. J. (1996). Simulation results for anchored clones. Technical Report 252, Department of Mathematics, Monash University, Australia.

Ewens, W. J., Bell, C. J., Donnelly, P. J., Dunn, P., Matallana, E. and Ecker, J. R. (1991). Genome mapping with anchored clones: Theoretical aspects. *Genomics* **11:** 799–805.

Hudson, T., Stein, L., Gerety, S., Ma, J., Castle, A., Silva, J., Slonim, D., Baptista, R., Kruglyak, L., Xu, S., Hu, X., Colbert, A., Rosenberg, C., Reeve-Daly, M., Rozen, S., Hui, L., Wu, X., Vestergaard, C., Wilson, K., Bae, J., Maitra, S., Ganiatsas, S., Evans, C., DeAngelis, M., Ingalls, K., Nahf, R., Horton, L., Oskin Anderson, M., Collymore, A., Ye, W., Kouyoumjian, V., Zemsteva, I., Tam, J., Devine, R., Courtney, D., Renaud, M., Nguyen, H., O'Connor, T., Fizames, C., Fauré, S., Gyapay, G., Dib, C., Morissette, J., Orlin, J., Birren, B., Goodman, N., Weissenbach, J., Hawkins, T., Foote, S., Page, D. and Lander, E. S. (1995). An STS-based map of the human genome. *Science* **270:** 1945–1954.

Nagaraja, R., MacMillan, S., Kere, J., Jones, C., Griffin, S., Schmatz, M., Terrell, J., Shomaker, M., Jermark, C., Hott, C., Masisi, M., Mumm, S., Srivastava, A., Pilia, T., G. and-Featherstone, Mazzarella, R., Kesterson, S., McCauley, B., Railey, B., Burough, F., Nowotny, V., D'Urso, M., States, D., Brownstein, B. and Schlessinger, D. (1997). X chromosome map at 75 kb STS resolution, revealing extremes of recombination and GC content. *Genome Research* **7:** 210–222.

Nelson, D. O. and Speed, T. P. (1994). Predicting progress in directed mapping projects. *Genomics* **24:** 41–52.

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.

Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. and Bernardi, G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene* **174:** 85–94.

Schbath, S. (1997). Coverage processes in physical mapping by anchoring random clones. *J. Comp. Biol.* **4:** 61–82.

Schbath, S. (1996). Using non-homogeneous processes in physical mapping by anchoring random clones: Mathematical analysis and application to hotspots. Technical report #96-6, Center for Applied Mathematical Sciences, University of Southern California, Los Angeles.

Schmidt, R., West, J., Cnops, G., Love, K., Balestrazzi, A. and Dean, C. (1996). Detailed description of four YAC contigs representing 17 Mb of chromosome 4 of *Arabidopsis thaliana* ecotype Columbia. *The Plant Journal* **9:** 755–765.