# Ordering Clone Libraries in Computational Biology

by

Martin Dyer
School of Computer Studies
University of Leeds, UK

Alan Frieze
Department of Mathematics
Carnegie Mellon University
Pittsburgh, PA, USA

Stephen Suen
Department of Mathematics
University of South Florida
Tampa, FL, USA

Research Report No. 94-172
October, 1994

# Ordering Clone Libraries in Computational Biology

Martin Dyer*
School of Computer Studies
University of Leeds, Leeds, UK

Alan Frieze†
Department of Mathematics
Carnegie Mellon University, Pittsburgh PA, USA

Stephen Suen
Department of Mathematics
University of South Florida, Tampa FL, USA

3 November, 1993

**Abstract**

We consider a probabilistic model, due to Lander and Waterman and to Alizadeh, Karp, Newberg and Weisser, for the physical mapping of DNA molecules. Within this model, we answer precisely a question of Alizadeh *et al* concerning the minimum number of probes required to reconstruct the entire ordering of a given clone library with high probability. We also examine the related problem of determining the least number of probes required to construct a "spine" for the library. We give a fairly precise characterization for this number.

*Supported in part by Esprit Working Group RAND.
†Supported in part by NSF grant CCR-9225008

# 1 Introduction

The objective of many efforts in molecular biology, including the Human Genome project, is to *sequence* chromosomal DNA, i.e. to obtain the sequence of A, C, T or G *nucleotides* which constitute one strand of each molecule. This may be a complex task since, for example, in a typical human chromosome there are on the order of $10^8$ nucleotides, and the whole *genome* comprises 23 pairs of chromosomes. Modelling this combinatorial complexity leads to interesting mathematical and computer science problems. See [6].

A sub-goal of DNA sequencing is often to construct a *physical map*, which specifies the location of specific identifiable fragments on the molecule. In the usual procedures for physical mapping, it is necessary to extract information from fragments of the molecule called *clones*. These might typically contain about $10^4$ nucleotides. A *clone library* is a collection of clones covering one or more molecules of interest, for example the human genome. One approach to physical mapping is to locate (by *hybridization*) the occurrence of short sequences called *probes* within the clones. See [2]. In practice, the identification process may also involve errors, which further complicates matters.

Lander and Waterman [4] proposed a probabilistic model for the location of clones in physical mapping. Their model was refined by Alizadeh, Karp, Newberg and Weisser [2] to encompass the occurrence of probes within clones. We will describe the model in detail in Section 2.

Assuming this model, the question was posed in [2] and [6] as to how many probes are required to correctly order a given clone library, even assuming that the data is error-free. ¿From the probabilistic viewpoint, the question concerns the existence of a "threshold" [1]. In Section 4 we will answer this question precisely. ¿From a practical viewpoint, the numbers required are fairly discouraging. Therefore, in Section 5, we will describe and examine a natural related problem, having a more modest objective than that of ordering the entire clone library. Again we can give reasonably precise characterizations for the numbers of probes required. For this second problem, in contrast with the first, the numbers required are rather more encouraging for the practitioner.

## 2 The model

Let $V = \{1, 2, \ldots, n\}$. Then we will consider the following model:

(a) The *genome* is represented by the interval $[0, L]$.

(b) There is a *library* of $n$ *clones*, each being an interval $C_i$ ($i \in V$) of unit length with $C_i \subset [0, L]$.

(c) The left end-points of the clones $X_i$ ($i \in V$) are independently and uniformly distributed in $[0, L-1]$. We assume that the $X_i$ are ordered in increasing order. (Note that with probability 1 no two $X_i$ are equal in this model.) We will say $C_i$ is "to the right" of $C_j$ if $X_i > X_j$, otherwise "to the left".

(d) There are $m$ *probes*, and the occurrences of each probe form a Poisson process with rate $\lambda$ on $[0, L]$. These $m$ Poisson processes are mutually independent.

We are interested in the case where $L$ is large, and order-of-magnitude statements will be as $L \to \infty$. Thus we will use the phrase "with high probability" in respect of any event to mean that its probability is $1 - o(1)$.

The problem is to correctly identify the ordering of the clones in the library using only the information contained in the probes. Thus for each clone and each probe we are told whether the clone contains the probe. The information can be represented as a $n \times m$ 0-1 matrix. (See, for example, [2] for more details.) We wish to determine the minimum number of probes which will allow us to correctly order the entire library. Assuming the above model, we answer this question precisely in Section 4, as follows. Let $p_0(n)$ be the probability that $n$ clones form one connected component, and $p_1(n)$ be the probability that there is a unique order for all the clones consistent with the probe information. Clearly $p_1(n) \leq p_0(n)$. Let us call $n(L)$ *minimal* if $p_0(n) \to 1$, but, for any $n'(L)$ such that $\lim_{L\to\infty} n'(L)/n(L) < 1$, we have $p_0(n') \to 0$. Then we show that

**Theorem 1** *Let $n(L)$ be minimal, and let $m = \beta(n) n \log n$ as $n \to \infty$, then*

*(i)* *If* $\beta(n) \to 0$, $p_1(n) \to 0$.

*(ii)* *If* $\beta(n) \to \beta$, *a constant,* $p_1(n) \to \exp\{-e^{\lambda}/(2\beta\lambda)\}$.

*(iii)* *If* $\beta(n) \to \infty$, $p_1(n) \to 1$.

Note that this number is in fact rather large, about $n \log n$. Therefore, in Section 5 we examine a more modest objective: that we correctly order a connected subset of the library which covers "almost all" of the genome. We call this a *spine* of the clone library.

**Remark** We will see that we need $n = L(\log L + \log\log L + \omega)$ where $\omega \to \infty$ with $n$. Our proof of Theorem 1 is given only for $\omega = o(\log L)$, but the reader may check that when $\omega = \gamma \log L$, Theorem 1 remains true with the exception that the limit in case (ii) is increased by a factor $(\gamma + 1)$.

# 3 Preliminaries

Let $X_0 = 0$, $X_{n+1} = L$ and $Z_i = X_i - X_{i-1}$ $(i = 1, \ldots, n+1)$. Suppose $W_i$ $(i = 1, \ldots, n+1)$ are independently and identically distributed unit exponentials. Then [5, p.75]

$$\frac{Z_i}{L} = \frac{W_i}{W_1 + \cdots + W_{n+1}} \qquad (i = 1, \ldots, n+1) \tag{1}$$

Thus we are interested in the quantity $T \doteq \sum_{j=1}^{n+1} W_j$. Now $T$ has probability density function $e^{-t}t^n/n!$ and

$$\Pr(T \geq t) = \sum_{j=0}^{n} t^j e^{-t}/j!.$$

Using this, we can show by standard estimations (see for example Alon and Spencer [1]) that

$$\Pr(|T/(n+1) - 1| > \epsilon) \leq 2\, e^{-\epsilon^2 n/4}/\epsilon \qquad (0 < \epsilon < 1).$$

Hence, for large $n$, $T/(n+1)$ will naturally tend to be very close to 1. For example, putting $\epsilon = 4\sqrt{\log n/n}$ gives

$$\Pr(|T/(n+1) - 1| > 4\sqrt{\log n/n}) = o(1/n^3),$$

and so

$$\Pr\left(\exists i : \left|\frac{LW_i}{(n+1)Z_i} - 1\right| > 4\sqrt{\frac{\log n}{n}}\right) = o(1/n^3) \qquad (i = 1, \ldots, n+1).$$

Hence with high probability all $Z_i$ can be approximated by $LW_i/n$ to relative error $O(\sqrt{\log n/n})$. We will use this approximation where appropriate.

Now observe that there is a "gap" in the coverage if $Z_i > 1$. However, letting $\epsilon = 4\sqrt{\log n/n}$,

$$\begin{aligned} \Pr(\forall i : Z_i \geq 1) &\leq \Pr(\forall i : W_i \leq (1+\epsilon)n/L) + o(1/n^3) \\ &= (1 - e^{-(1+\epsilon)n/L})^n + o(1/n^3). \end{aligned}$$

Now suppose $n = L(\log L + \log\log L + \omega)$ for some $\omega$, we have

$$\Pr(\forall i : Z_i \leq 1) \leq \exp(-(1 - o(1))e^{-\omega}) + o(1/L^3).$$

Thus, if $\omega \to -\infty$, with high probability there will exist a gap. On the other hand, an almost identical calculation gives

$$\Pr(\forall i : Z_i \leq 1) \geq \exp(-(1 + o(1))e^{-\omega}) - o(1/L^3),$$

so if $\omega \to \infty$, with high probability there will exist no gap. Hence constant $\omega$ is the threshold for connectivity of the clones. This result is well known in different contexts, see [3] and the references contained therein. Note that ordering will be impossible if the clone library is not connected since, no matter how many probes are used, there will be know way of detecting the order in which the disconnected pieces should be placed. Thus, to ensure connectivity with high probability, we must have $\omega \to \infty$. Note now that, in our earlier definition, the clone library is minimal if and only if $\omega = o(\log n)$. However, most of our proofs are easily modifed for the case $\omega = \Omega(\log n)$. The

details are left to the reader. However, in practice, clone libraries are designed to provide small constant coverage of the genome, with a coverage factor of about five, say. In the model, $k$-times coverage of an interval including most of the genome corresponds to taking $n(L)$ at around $L(\log L + k \log\log L)$, or $\omega$ at about $(k-1)\log\log L$. (See, by comparison, [3].) Thus this region of greatest interest for $\omega$ falls well within the scope of our results.

It will be observed, that in this model, short segments at either end of the genome will (with probability 1) be uncovered by any clone. This is a small deficiency in the model which we ignore, assuming that these segments are to be handled separately.

# 4 Threshold

Let $m = \beta n \log n$. We show that the threshold for sequencing the entire set of clones occurs at constant $\beta$, and hence prove Theorem 1.

Let $I_i$ denote the set of probes which fall in clone $C_i$ ($i \in V$). Consider the collection $\mathcal{D}_i$ of all difference sets $\Delta_{j,i} = I_j \setminus I_i$ for $i, j \in V$ and $i \neq j$, and let $D_{j,i} = |\Delta_{j,i}|$. We show first that with high probability adjacent clones have smaller difference sets than disjoint clones. We prove this is a slightly stronger form than is needed here, for use in Section 5.

**Lemma 1** *Let $\delta$ be such that (i) $\delta = o(\omega/\log n)$ and (ii) $\sqrt{(\log n)/n} = o(\delta)$. Let $D^*(\delta) = me^{-\lambda}(1-e^{-\lambda})(1-\delta)$, and $m(n)$ be such that $(\log n)/\delta^2 = o(m)$ as $n \to \infty$. Then, for every $i \in V$, with high probability*

$$D_{i+1,i} < D^* \text{ and } \min\{D_{j,i} : C_j \cap C_i = \emptyset\} > D^*.$$

**Proof** Let $\delta'$ be such that $\delta = o(\delta')$ and $\delta' = o(\omega/\log n)$. Then

$$\begin{aligned}
\Pr(\forall i : Z_i \leq 1-\delta') &\geq \Pr(\forall i : W_i \leq (1-2\delta')n/L) - o(1/n^3) \\
&\geq (1 - e^{-(1-3\delta')(\log n+\omega)})^n + o(1/n^3) \\
&\geq \exp(-e^{-w/2}n^{4\delta'}) + o(1/n^3) \\
&= 1 - o(1),
\end{aligned}$$

for large $L$. Thus, assume that all $Z_i < 1 - \delta'$. Now, if $C_j \cap C_i = \emptyset$, then probes fall in $I_j \setminus I_i$ independently Hence we have

$$\Pr(\exists i, j : D_{j,i} < D^*) \leq n^2 \exp\{-\tfrac{1}{3}\delta^2 m e^{-\lambda}(1 - e^{-\lambda})\} = o(1).$$

On the other hand, probes fall in $D_{i+1,i}$ independently with probability dominated by

$$e^{-\lambda}(1 - e^{-\lambda(1-\delta')}) \approx \left(1 - \frac{\lambda}{e^{\lambda} - 1}\delta'\right) e^{-\lambda}(1 - e^{-\lambda}).$$

But, for large enough $n$, $(1 - \lambda\delta'/(e^{\lambda} - 1))(1 + \delta) < (1 - \delta)$. Thus

$$\Pr(D_{i+1,i} > D^*) \leq n \exp\{-(\tfrac{1}{3} - o(1))\delta^2 m e^{-\lambda}(1 - e^{-\lambda})\} = o(1).$$

□

Thus, all information concerning the ordering of the clones is contained in the "small" sets in the $\mathcal{D}_i$. Let us define a *minimal* set in $\mathcal{D}_i$ to be one which is not *properly* contained in any other set in $\mathcal{D}_i$. Let

$$\mathcal{M}_i = \{\Delta \in \mathcal{D}_i : \Delta \text{ is minimal and } |\Delta| < D^*\}.$$

Now consider the graph $G = (V, E)$, where $\{i, j\} \in E$ if either $\Delta_{j,i} \in \mathcal{M}_i$ or $\Delta_{i,j} \in \mathcal{M}_j$. We now prove the crucial properties of $G$.

**Lemma 2** *With high probability $\{i, k\}$ is not an edge for all $|k - i| > 2$.*

**Proof** Suppose the Lemma is false, and let us assume without loss that $k > i+2$ and $\Delta_{k,i} \in \mathcal{M}_i$. (The case $k < i-2$ is symmetric.) Hence $C_k \cap C_i \neq \emptyset$ by Lemma 1. Suppose $i < j < k$. Then $C_j \subseteq C_k \cup C_i$. Hence $I_j \subseteq I_k \cup I_i$. Thus $I_j \setminus I_i \subseteq I_k \setminus I_i$, i.e. $\Delta_{j,i} \subseteq \Delta_{k,i}$ and we must have $\Delta_{j,i} = \Delta_{k,i}$ by minimality. Hence $\Delta_{r,i} = \Delta_{i+1,i}$ for all $i < r \leq k$. We have four successively overlapping intervals $C_i, C_{i+1}, C_{i+2}, C_{i+3}$. The expected total number of quadruples of successively overlapping intervals is at most

$$\frac{n(n-1)(n-2)(n-3)}{L^3} < \frac{n^4}{L^3}.$$

Given such a quadruple, let $w = Z_{i+1}, x = Z_{i+2}, y = Z_{i+3}$. If the quadruple satisfies $\Delta_{i+1,i} = \Delta_{i+2,i} = \Delta_{i+3,i}$, then every probe which falls in $[X_{i+1} + 1, X_{i+3} + 1]$ must also fall in $[X_i, X_{i+1} + 1]$. This has probability

$$(1 - e^{-\lambda(1+w)}(1 - e^{-\lambda(x+y)}))^m.$$

Thus the event $\mathcal{E}_1$ that there exists any quadruple meeting this condition satisfies

$$\begin{aligned}
\Pr(\mathcal{E}_1) &< \frac{n^4}{L^3}\int_0^1 dw \int_0^1 dx \int_0^1 dy (1 - e^{-\lambda(1+w)}(1 - e^{-\lambda(x+y)}))^m \\
&= \frac{n^4}{L^3}\int_0^1 dw \int_0^a dx \int_0^a dy (1 - \tfrac{1}{2}\lambda e^{-2\lambda}(x+y))^m + o(n^{-1}) \\
&\qquad\qquad \text{where } a = n^{-1/2}, \\
&\leq \frac{n^4}{L^3}\int_0^1 dw \int_0^1 dx \int_0^1 dy (1 - \tfrac{1}{2}\lambda e^{-2\lambda}(x+y))^m \\
&\leq \frac{n^4}{L^3}\int_0^1 dw \left(\int_0^a e^{-bmz} dz\right)^2 \qquad \text{where } b = \lambda e^{-2\lambda}/2, \\
&= \frac{n^4}{L^3} \times O\left(\frac{1}{m^2}\right) \\
&= O\left(\frac{\log n}{n}\right).
\end{aligned}$$

□

**Lemma 3** *With high probability* $\{i, i+1\}$ *is an edge for* $i = 1, 2, \ldots, n-1$.

**Proof** Otherwise there exists an $i$ such that $\Delta_{i+1,i} \notin \mathcal{M}_i$ and $\Delta_{i,i+1} \notin \mathcal{M}_{i+1}$. This only occurs if, for some $i$, $\Delta_{i-1,i} \subset \Delta_{i+1,i}$ and $\Delta_{i+2,i+1} \subset \Delta_{i,i+1}$. Let us call this event $\mathcal{E}_2$, and let $w = Z_i, x = Z_{i+1}, y = Z_{i+2}$. As in Lemma 2, we have a quadruple of sets which overlap in succession, and if $\mathcal{E}_2$ occurs, every probe which falls in either $[X_{i-1}, X_i]$ or $[X_{i+1} + 1, X_{i+2} + 1]$ must also fall in $[X_i, X_{i+1} + 1]$. This has probability $(1 - e^{-\lambda(1+x)}(1 - e^{-\lambda(w+y)}))^m$. Hence, similarly to Lemma 2,

$$\Pr(\mathcal{E}_2) < \frac{n^4}{L^3}\int_0^1 dw \int_0^1 dx \int_0^1 dy (1 - e^{-\lambda(1+x)}(1 - e^{-\lambda(x+y)}))^m = O\left(\frac{\log n}{n}\right).$$

□

From Lemmas 2 and 3, we see that $G$ will be a path with the possible exception of cases where an edge of the form $\{i, i+2\}$ exists. If such an edge exists then, unless $\{i-1, i+1\}$ is also present, $G$ decomposes into two smaller graphs joined by the bridge $\{i-1, i\}$. Hence we can recursively subdivide the order-reconstruction problem. The only difficulty occurs when both $\{i-1, i+1\}$ and $\{i, i+2\}$ are present, since then it is not clear whether $i+1$ should be placed before or after $i$. Let $\mathcal{E}_3$ be the event that this happens for some $i$, let $\mathcal{E}_4^{(i)}$ be the event $\{I_{i+1} = I_i\}$ and $\mathcal{E}_4 = \bigcup_{i=1}^{n} \mathcal{E}_4^{(i)}$. Note that $\mathcal{E}_4^{(i)} = \{\Delta_{i+1,i} = \Delta_{i,i+1} = \emptyset\}$. Clearly, if $\mathcal{E}_4$ occurs, we cannot order all the clones, since there will be a pair $C_i, C_{i+1}$ which could be ordered incorrectly but consistently with the data. On the other hand, if $\mathcal{E}_3$ does not occur, then we have argued that ordering is possible. (And is, in fact, achievable by a simple polynomial time algorithm). Clearly $\mathcal{E}_4$ implies $\mathcal{E}_3$, since if $I_{i+1} = I_i$ then $\Delta_{i,i-1} = \Delta_{i+1,i-1}$ and $\Delta_{i,i+2} = \Delta_{i+1,i+2}$. We now show that the converse is almost always true.

**Lemma 4** $\Pr(\mathcal{E}_3 \setminus \mathcal{E}_4) = o(1)$.

**Proof** We will use the notation from the proof of Lemma 3. If $\mathcal{E}_3$ occurs, then for some $i$, the following two events occur.

(i) Either $\Delta_{i-1,i+1} \in \mathcal{M}_{i+1}$ or $\Delta_{i+1,i-1} \in \mathcal{M}_{i-1}$,

(ii) Either $\Delta_{i+2,i} \in \mathcal{M}_i$ or $\Delta_{i,i+2} \in \mathcal{M}_{i+2}$.

Reasoning as in Lemmas 2 and 3, this gives four possible cases:

(a) For some $i$, $\Delta_{i-1,i+1} = \Delta_{i,i+1}$ and $\Delta_{i+2,i} = \Delta_{i+1,i}$. Call this event $\mathcal{E}_a$. we see that every probe which falls in either $[X_{i-1}, X_i]$ or $[X_{i+1}+1, X_{i+2}+1]$ must also fall in $[X_i, X_{i+1}+1]$. The probability of this was bounded by $O(\log n/n)$ in the proof of Lemma 3. Thus $\Pr(\mathcal{E}_a) = O(\log n/n)$.

(b) For some $i$, $\Delta_{i-1,i+1} = \Delta_{i,i+1}$ and $\Delta_{i,i+2} = \Delta_{i+1,i+2}$. Call this event $\mathcal{E}_b$. We see that every probe which falls in $[X_{i-1}, X_i]$ falls in $[X_i, X_{i+1}+1]$ and every probe which falls in $[X_i, X_{i+1}]$ falls in $[X_{i+1}, X_{i+2}+1]$. Thus every probe which falls in $[X_{i-1}, X_{i+1}]$ falls in $[X_{i+1}, X_{i+2}+1]$. But the probability

of this event can bounded as in the proof of Lemma 2. Thus

$$\Pr(\mathcal{E}_b) < \frac{n^4}{L^3}\int_0^1 dw \int_0^1 dx \int_0^1 dy (1-e^{-\lambda(1+y)}(1-e^{-\lambda(w+x)}))^m = O\left(\frac{\log n}{n}\right).$$

(c) For some $i$, $\Delta_{i+1,i-1} = \Delta_{i,i-1}$ and $\Delta_{i+2,i} = \Delta_{i+1,i}$. This event, $\mathcal{E}_c$ say, is simply the reverse (in clone ordering) of $\mathcal{E}_b$, and hence $\Pr(\mathcal{E}_c) = O(\log n/n)$.

(d) For some $i$, $\Delta_{i+1,i-1} = \Delta_{i,i-1}$ and $\Delta_{i,i+2} = \Delta_{i+1,i+2}$. Call this event $\mathcal{E}_d^{(i)}$, for given $i$, and let $\mathcal{E}_d = \bigcup_{i=1}^n \mathcal{E}_d^{(i)}$. Note that $\mathcal{E}_4^{(i)} \subseteq \mathcal{E}_d^{(i)}$, so we must estimate more carefully than above. If $\mathcal{E}_d^{(i)}$ occurs, then every probe which falls in $[X_i+1, X_{i+1}+1]$ falls in $[X_{i-1}, X_i+1]$ and every probe which falls in $[X_i, X_{i+1}]$ falls in $[X_{i+1}, X_{i+2}+1]$. The probability of this event is

$$\left(1-(1-e^{-\lambda x})(e^{-\lambda(1+w)}+e^{-\lambda(1+y)})\right)^m,$$

since the following two events are disjoint. A given probe falls

(i) in $[X_i+1, X_{i+1}+1]$ and not in $[X_{i-1}, X_i+1]$.

(ii) in $[X_i, X_{i+1}]$ and not in $[X_{i+1}, X_{i+2}+1]$.

Let $N_d$ be the number of events $\mathcal{E}_d^{(i)}$ which occur. Then, using the exponential approximation to the distribution of the $Z_j$ $(j = i, i+1, i+2)$, and writing $\alpha = n/L = \log L + \log\log L + \omega$,

$$\begin{aligned}
&\mathbf{E}(N_d) \approx \\
&n\int_0^\infty\int_0^\infty\int_0^\infty \left(1-(1-e^{-\lambda x})(e^{-\lambda(1+w)}+e^{-\lambda(1+y)})\right)^m \alpha^3 e^{-\alpha(w+x+y)}\,dw\,dx\,dy \\
&\quad\approx\ n\int_0^\infty\int_0^\infty\int_0^\infty \exp\left(-m\lambda e^{-\lambda}(e^{-\lambda w}+e^{-\lambda y})x\right)\alpha^3 e^{-\alpha(w+x+y)}\,dw\,dx\,dy \\
&\quad=\ n\alpha\int_0^\infty\int_0^\infty \frac{\alpha^2 e^{-\alpha(w+y)}}{m\lambda e^{-\lambda}(e^{-\lambda w}+e^{-\lambda y})+\alpha}\,dw\,dy \\
&\quad\approx\ \frac{n\alpha}{m\lambda e^{-\lambda}}\int_0^\infty\int_0^\infty \frac{\alpha^2 e^{-\alpha(w+y)}}{e^{-\lambda w}+e^{-\lambda y}}\,dw\,dy \\
&\quad=\ \frac{n\alpha}{m\lambda e^{-\lambda}}\int_0^\infty\int_0^\infty \frac{e^{-(w+y)}}{e^{-\lambda w/\alpha}+e^{-\lambda y/\alpha}}\,dw\,dy
\end{aligned}$$

$$\approx \frac{n\alpha}{m\lambda e^{-\lambda}} \cdot \frac{1}{2} \qquad \text{(since } \alpha \to \infty \text{ with } n\text{)}$$

$$= \frac{e^{\lambda}}{2\beta\lambda} \qquad \text{(since } \alpha \approx \log n,\ m = \beta n \log n\text{)}.$$

Now let $N$ be the number of events $\mathcal{E}_i$ which occur. The event $\mathcal{E}_4^{(i)}$ occurs if and only if there is no probe in $[X_i, X_{i+1}]$ which is not in $[X_{i+1}, X_{i+1}+1]$ and no probe in $[X_i+1, X_{i+1}+1]$ which is not in $[X_i, X_i+1]$. Thus, using disjointness as in (d) above,

$$\begin{aligned}
\mathbf{E}(N) &\approx n\int_0^{\infty}(1-2e^{-\lambda}(1-e^{-\lambda z})^m \alpha e^{-\lambda z}dz \\
&\approx n\int_0^{\infty}\exp(2m\lambda e^{-\lambda}z)\alpha e^{-\lambda z}dz \\
&\approx \frac{n\alpha}{2m\lambda e^{-\lambda}} \\
&\approx \frac{e^{\lambda}}{2\beta\lambda}.
\end{aligned}$$

But this is asymptotically equal to $\mathbf{E}(N_d)$. Now $\mathcal{E}_4^{(i)} \subseteq \mathcal{E}_3^{(i)}$, and hence

$$\begin{aligned}
\Pr(\mathcal{E}_3 \cap \bar{\mathcal{E}}_4) &\leq \Pr\left(\bigcup_{i=1}^{n}\left(\mathcal{E}_3^{(i)} \cap \bar{\mathcal{E}}_4{}^{(i)}\right)\right) \\
&\leq \sum_{i=1}^{n}\Pr(\mathcal{E}_3^{(i)} \cap \bar{\mathcal{E}}_4{}^{(i)}) \\
&= \sum_{i=1}^{n}\{\Pr(\mathcal{E}_3^{(i)}) - \Pr(\mathcal{E}_4^{(i)})\} \\
&= \mathbf{E}(N_d) - \mathbf{E}(N) \\
&= o(1).
\end{aligned}$$

We have therefore shown that, asymptotically, ordering depends on the occurence or otherwise of $\mathcal{E}_4$, and that the expected number of such events is $e^{\lambda}/(2\beta\lambda)$. We now complete the proof of Theorem 1. □

**Lemma 5** $\Pr(\bar{\mathcal{E}}_4) \to \exp\{-e^{\lambda}/(2\beta\lambda)\}$ *as* $n \to \infty$.

**Proof** Let $\mu = \mathbf{E}(N) = e^{\lambda}/(2\beta\lambda)$. Then $\Pr(\mathcal{E}_i) \approx \mu/n$. Let us say $C_j$ *misses* $C_i$ if $C_j \cup C_{j+1}$ does not meet $C_i \cup C_{i+1}$. Let $\nu_i$ be the number of $C_j$ which do not miss a given $C_i$. Then if $t = \lceil 10 \log n \rceil$, say,

$$\Pr(\max_i \nu_i \geq 40 \log n) \leq 4\binom{n}{t}\frac{1}{L^t} < n^{-10}.$$

Thus we may assume $\max_i \nu_i < 40 \log n$. But, for given $i$ and $j$ such that $C_i$ does not miss $C_j$, a calculation similar to that in part (a) of the proof of Lemma 4 gives $\Pr(\mathcal{E}_i \cap \mathcal{E}_j) = O(1/m^2)$. Thus $\Pr(\mathcal{E}_j \mid \mathcal{E}_i) = O(n/m^2)$ However, if $C_i$ misses $C_j$ then $\mathcal{E}_i, \mathcal{E}_j$ are independent. Consider $\sum_S \Pr(\bigcap_{i \in S} \mathcal{E}_i)$ over all sets $S$ size $k$ for constant $k$. If $C_i$ misses $C_j$ for all $i, j \in S$ then $\Pr(\bigcap_{i \in S} \mathcal{E}_i) \approx (\mu/n)^k$. Hence the contribution to the sum from these terms is $\mu^k/k!$, since there are about $n^k/k!$ such sets. However if $S$ contains two sets that do not miss, let $r$ be the maximum number of $i \in S$ such that the $C_i$ miss one another. The contribution from such sets is then at most

$$\sum_{r=1}^{k-1} \frac{n^r}{r!}(40r \log n)^{k-r} \left(\frac{\mu}{n}\right)^r \times O\left(\frac{n}{m^2}\right) = O\left(\frac{(\log n)^{k-3}}{n}\right).$$

Hence, for $k = 1, 2, 3, \ldots$, we have

$$\sum_{S:|S|=k} \Pr(\bigcap_{i \in S} \mathcal{E}_i) \approx \frac{\mu^k}{k!}.$$

Then, by inclusion-exclusion,

$$\Pr(\bigcup_{i \in V} \mathcal{E}_i) \to \sum_{k=1}^{\infty} (-1)^{k-1} \frac{\mu^k}{k!} = 1 - e^{-\mu}.$$

□

Hence, we have Theorem 1 for constant $\beta$. For $\beta \to \infty$ slowly enough, we therefore have $\Pr(\mathcal{E}_4) \to 1$, and for $\beta \to 0$ slowly enough, we have $\Pr(\mathcal{E}_4) \to 0$. Theorem 1 now follows by observing that the probability that we can reconstruct the clone ordering is monotone in the number of probes for a fixed number of clones.

# 5 Constructing a spine

In this section, we consider the problem of constructing a *spine*, i.e. a *subset* of the clone library which is correctly ordered and covers the genome, with the exception only of regions of length $o(1)$ at either end as $L \to \infty$. We will show that this requires considerably fewer probes than in Section 4, since we are not obliged to give any ordering information on the clones which are not in the spine. We will show that $m = o(\log^3 L)$ probes will always suffice, and that $\Theta(\log L)$ probes are both necessary and sufficient for this task.

Suppose $C_i \cap C_j \neq \emptyset$, and $X_j - X_i = x$. Then $D_{j,i}$ is binomial with parameters $m$, $e^{-\lambda}(1-e^{-\lambda x})$. Let us write $\mu(x) = me^{-\lambda}(1-e^{-\lambda x})$ for the expected value. Clearly $\mu(x)$ is an increasing function of $x$.

Let $\delta = o(\delta')$ and $\delta' = o(\omega/\log n)$ as in Lemma 1, and also let $\epsilon$ be such that $\epsilon = o(\delta')$, and $\delta = o(\epsilon)$. Then, for each $i \in V$, define

$$S_i = \{j : \mu(\epsilon^2) \leq D_{j,i} \leq \mu(1-\epsilon)\}.$$

Now consider the following algorithm for finding an ordered subset of the clones.

(0) Pick any clone $C_{i_1}$, and find a $j_1 \in S_i$ such that $D_{j_1,i_1}$ is maximum. Place clone $C_{j_1}$ adjacent to $C_{i_1}$ in the chosen subset. Let $i \leftarrow i_1, j \leftarrow j_1$.

(1) If there is $k \in S_j$ such that $D_{i,k} > D_{j,k}$ and $D_{k,i} > D_{j,i}$, pick one such that $D_{k,j}$ is maximum. Place clone $C_k$ adjacent to $C_j$ on the opposite side from $C_i$ in the chosen subset. Let $i \leftarrow j$, $j \leftarrow k$ and repeat this step.

(2) Otherwise reverse direction. Let $j \leftarrow i_1$, $i \leftarrow j_1$. Repeat step (1) until the loop terminates again, then stop.

**Theorem 2** *If $m(n)$ is such that $\log n/\delta^2 = o(m)$, then with high probability the algorithm will succeed in correctly determining a linear order on a subset of the clones. Futhermore, if $S$ is the set of clone numbers in the chosen*

*subset, then* $|S| = O(L)$ *and*

$$\min_{j \in S} X_j - X_1 < \delta^2, \quad X_n - \max_{j \in S} X_j < \delta^2.$$

We will prove the theorem in the following sequence of Lemmas.

**Lemma 6** *If* $j \in S_i$, *then* $C_i \cap C_j \neq \emptyset$.

**Proof** This follows directly from Lemma 1. □

**Lemma 7** *If* $X_n - X_i \geq \delta^2$, *then* $S_i \neq \emptyset$.

**Proof** Note that $\mu(1-\epsilon) > D^*(\delta')$, as in Lemma 1. Thus, if $i \neq n$, there is a least one $j$ (i.e. $i+1$) such that $D_{j,i} \leq \mu(1-\epsilon)$. Also, $\mu(\epsilon^2) \approx m\lambda e^{-\lambda}\epsilon^2 < \frac{1}{2}m\lambda e^{-\lambda}\delta'^2 \approx \frac{1}{2}\mu(\delta'^2)$. Thus, if $X_j - X_i > \delta'^2$ for any $i, j$, then

$$\Pr(D_{j,i} < \mu(\epsilon^2)) \leq \exp(-\tfrac{1}{3}(\tfrac{1}{2})^2\mu(\delta'^2)).$$

Hence

$$\Pr(\exists i, j : X_j - X_i > \delta'^2 \text{ and } D_{j,i} < \mu(\epsilon^2)) \leq n^2 \exp(-\tfrac{1}{3}(\tfrac{1}{2})^2\mu(\delta'^2)) = o(1),$$

since $\log n = o(\mu(\delta'^2))$.

Thus if $S_i = \emptyset$, we may assume $X_j - X_i < \delta'^2$ for all $j$ such that $C_i \cap C_j \neq \emptyset$. Let $X_j$ be maximum such that $X_j \leq X_i + 1$. Clearly $X_{j+1} \notin [X_j, X_i + 1]$, and thus if $j < n$ we have $Z_{j+1} > 1 - \delta'^2$, contradicting the proof of Lemma 1. □

**Lemma 8** *If* $X_n - X_j \geq \delta^2$, *there will exist a choice for* $k$ *in step (1) of the algorithm.*

**Proof** We have $\mu(\epsilon^2) \approx m\lambda e^{-\lambda}\epsilon^2 > 2m\lambda e^{-\lambda}\delta^2 \approx 2\mu(\delta^2)$. If $X_j - X_i < \delta^2$ for any $i$, $j$ with $j \in S_i$, then

$$\Pr(D_{j,i} \geq \mu(\epsilon^2)) \leq \exp(-\tfrac{1}{3}\mu(\delta^2)).$$

Hence,

$$\Pr(\exists i,j : X_j - X_i < \delta^2 \text{ and } D_{j,i} \geq \mu(\epsilon^2)) \leq n^2 \exp(-\tfrac{1}{3}\mu(\delta^2)) = o(1),$$

since since $\log n = o(\mu(\delta^2))$. Thus we may assume that $X_j - X_i > \delta^2$. If $j \neq n$, from Lemma 7 there will exist a $k \in S_j$, and we will have $X_k - X_j > \delta^2$.

First suppose $C_i \cap C_k \neq \emptyset$, so $C_i \cap C_j \cap C_k \neq \emptyset$. Then $\Delta_{j,k} \subseteq \Delta_{i,k}$, thus $D_{j,k} \geq D_{i,k}$ only if $\Delta_{j,k} = \Delta_{i,k}$. But since $X_j - X_i > \delta^2$ and $X_k - X_j < 1$, the existence of such an $i$, $j$ and $k$ has probability less than

$$n^3(1 - e^{-2\lambda}(1 - e^{-\lambda\delta^2}))^m = o(1).$$

Similarly, since $X_k - X_j > \delta^2$ and $X_j - X_i < 1$, $n^3 \Pr(D_{j,i} \geq D_{k,i}) = o(1)$.

Suppose finally, that $C_i \cap C_k = \emptyset$. Now $D_{j,i} < \mu(1-\epsilon) < D^*(\delta)$, since $j \in S_i$. But, from Lemma 1, $D_{k,i} > D^*(\delta)$, so $D_{k,i} > D_{j,i}$. Similarly $D_{i,k} > D^*(\delta)$. Let $\epsilon'$ be such that $\epsilon' = o(\epsilon)$ and $\delta = o(\epsilon')$. Then $\mu(1-\epsilon) < (1-\epsilon')\mu(1-\epsilon')$, so if $X_k - X_j > 1 - \epsilon'$, then

$$\Pr(D_{k,j} < \mu(1-\epsilon)) < \exp(-\tfrac{1}{3}\epsilon'^2\mu(1-\epsilon')) = o(n^{-2}).$$

Inflating the right side of the above inequality by $n^2$ deals with the existence of any such pair $k, j$. Since $D_{k,i} \geq D_{k,j}$ we may now assume $X_k - X_j \leq 1 - \epsilon'$. But $D^*(\delta) \geq (1+\epsilon')\mu(1-\epsilon')$, so

$$\Pr(D_{j,k} > D^*(\delta))) < \exp(-\tfrac{1}{3}\epsilon'^2\mu(1-\epsilon')) = o(n^{-2}),$$

and inflation by $n^2$ can be used as previously. □

**Lemma 9** *It is correct to place $C_k$ on the opposite side of $C_j$ from $C_i$.*

**Proof** Suppose $X_k < X_i < X_j$. Then, since $k \in S_j$, we have $C_k \cap C_j \neq \emptyset$ and hence $C_i \cap C_j \cap C_k \neq \emptyset$. But then $D_{j,k} \geq D_{i,k}$, a contradiction.

Thus suppose $X_i < X_k < X_j$. Then, since $j \in S_i$, we have $C_i \cap C_j \neq \emptyset$ and hence $C_i \cap C_j \cap C_k \neq \emptyset$. But then $D_{j,i} \geq D_{k,i}$, again a contradiction. □

Theorem 2 now follows. We see that when we are building the subset from "left to right", we cannot terminate until we have encountered a $C_j$ such that

$X_n - X_j < \delta^2$. Similarly, by symmetry, when we are building the subset from right to left, we cannot terminate until we have encountered a $C_j$ such that $X_j - X_1 < \delta^2$. Thus we have only to verify the claim concerning the number of clones selected.

**Lemma 10** *The algorithm will select $O(L)$ clones.*

**Proof** We will merely sketch the method, leaving the routine calculations to the reader.

Let $j \in B$ if and only if there is no $k$ for which $X_j + \frac{1}{3} < X_k < X_j + \frac{2}{3}$. Then, since $\Pr(X_i > L - 2) = 1/(L-1)$,

$$\mathbf{E}(|B|) \le n\left(1 - \frac{1}{3L}\right)^n + O(1/L) \le n\exp(-\tfrac{1}{3}(\omega + \log n)) + O(1/L) = o(n^{2/3}).$$

Thus $\Pr(|B| > n^{2/3}) = o(1)$. If $j \notin B$, let $k$ be such that $X_j + \frac{1}{3} < X_k < X_j + \frac{2}{3}$. By straightforward calculations we obtain

$$\Pr(D_{k,j} < \mu(\epsilon^2)) = o(n^{-2}), \quad \Pr(D_{k,j} > \mu(1-\epsilon)) = o(n^{-2}).$$

Thus with high probability we will have $k \in S_j$ in all such cases.

Now, if $X_j + \frac{1}{3} < X_k < X_j + \frac{2}{3}$ and $X_i < X_j$, similar calculations give

$$\Pr(D_{i,k} \le D_{j,k}) = o(n^{-3}), \quad \Pr(D_{k,i} \le D_{j,i}) = o(n^{-3}).$$

Thus $k$ will be a valid selection in step (1).

However, if $\ell \in S_j$ is such that $X_\ell < X_j + \frac{1}{4}$, then letting $t = \frac{1}{2}(\mu(\frac{1}{4}) + \mu(\frac{1}{3}))$, further calculations show

$$\Pr(D_{\ell,j} \ge t) = o(n^{-2}), \quad \Pr(D_{k,j} \le t) = o(n^{-2}),$$

so with high probability $D_{\ell,j} < D_{k,j}$. Thus if $j \notin B$, the maximization in step (1) will ensure that we choose $k$ so that $X_k - X_j \ge \frac{1}{4}$.

Now suppose the algorithm chooses $g$ clones $C_i$ with $i \notin B$, and $b$ with $i \in B$. Then these cover an interval of length at least $\frac{1}{4}(g-1) \le L$, so $g \le 4L + 1$. Thus the total number selected is

$$g + b \le 4L + 1 + n^{2/3} = O(L).$$

□

Note that if $m(n) = \omega' \log n/\delta^2$ and $\delta = \omega'/\log n$, where $\omega' = o(\omega)$ but $\omega' \to \infty$, then

$$m = \log^3 n/\omega' = o(\log^3 n).$$

Thus $o(\log^3 n)$ clones will always suffice. On the other hand, suppose $\omega = \log n/\sqrt{\omega'}$, where $\omega'$ tends arbitrarily slowly to infinity. Thus the the number of clones is "maximally minimal". Then letting $\delta = 1/\omega'$, say, gives $m = (\omega')^3 \log n$, so we need a little more than $\log n$ probes.

We also have the following

**Corollary 1** *If $m(n) = \log^3 n$, then $\bigcup_{j \in S} C_j = \bigcup_{i=1}^{n} C_i$.*

**Proof** First suppose $\omega = O(\sqrt{\log n})$. Take $\delta = \omega'/\log n$, where $\omega' = o(\omega)$ but $\omega' \to \infty$. Thus $\log^3 n$ probes suffice, and $\delta^2 = o(1/\log n)$. Hence

$$\Pr(Z_n < \delta^2) = 1 - \exp(-(1 + o(1))\delta^2 \log n) = o(1),$$

and thus, by Lemma 8, the loop in algorithm must terminate with $j = n$. Similarly for the reverse direction.

Now if $\sqrt{\log n} = o(\omega)$, take $\delta = 1/\omega$. Then a little greater than $\omega^2 \log n = o(\log^3 n)$ probes suffice, and again $\delta^2 = o(1/\log n)$. The remainder of the argument is as before. □

Thus, with $\log^3 n$ probes, we guarantee to cover the entire interval represented in the clone library with high probability .

Although our proofs are given only for $\omega = o(\log L)$, similar methods extend to faster growing values of $\omega$. However, we may deduce from the above that, if we take $\omega = \Theta(\log L)$, $O(\log n)$ probes will suffice. We use the following simple Lemma.

**Lemma 11** *Let $f(n), g(n)$ be positive functions. If $f = o(h)$ for all positive $h(n)$ such that $g = o(h)$ as $n \to \infty$, then $f = O(g)$.*

**Proof** If $f \neq O(g)$, there is an increasing sequence $c_i \to \infty$ such that for all $i$, there exists $n_i$ such that $f(n_i) > c_i g(n_i)$. Assume without loss that $n_i$

is nondecreasing, and define $h$ by $h(n) = c_i g(n)$ if $n_i \leq n < n_{i+1}$. Clearly $g = o(h)$, so $f = o(h)$. Thus $f(n_i) < c_i g(n_i)$ for large $i$, a contradiction. □

**Lemma 12** *Let $\varepsilon > 0$. If $n(L) \geq (1+\varepsilon) \log L$, then $O(\log n)$ probes suffice to determine the spine.*

**Proof** For a fixed number of probes $m$, the algorithm above clearly cannot have smaller probability of success with a larger number of clones. Thus the number $m(n)$ required for any given probability of success is nondecreasing with $n$. Thus, if $m^*$ is the number required when $n = (1+\varepsilon)L \log L$, $m^* \leq m(n)$ for all $n$ such that $\omega = o(\log n)$. Let $h$ be arbitrary such that $\log n = o(h)$, and let $\tau = (h/\log n)^{1/6}$. Thus $\tau \to \infty$. Let $\omega = \log n/\tau$, $\delta = 1/\tau^2$, and $m = \tau^5 \log n$. Then Theorem 2 applies, and hence $m^* \leq m = o(h)$. We now apply Lemma 11, with $g = \log n$, $f = m^*$ to complete the proof. □

On the other hand, we have the following simple lower bound.

**Lemma 13** *At least $L \log L$ clones and $\log_2 L$ probes are necessary to determine a spine.*

**Proof** We need $L \log L$ clones for connectedness, without which we clearly cannot determine a spine. Any spine must obviously have at least $L$ clones. Since it is unambiguously ordered, no two clones can contain the same set of probes. Now, with $m$ probes, there are at most $2^m$ different sets. Thus $2^m \geq L$. □

We may sumarise these results in the following "optimality theorem".

**Theorem 3** *$n = \Theta(L \log L)$ clones and $m = \Theta(\log L)$ probes are necessary and sufficient to determine a spine.*

# References


[1] N. Alon, J. Spencer and P. Erdős, *The Probabilistic Method*, John Wiley and Sons, New York, 1992.

[2] F. Alizadeh, R. M. Karp, L. A. Newberg and D. K. Weisser, "Physical mapping of chromosomes: A combinatorial problem in molecular biology", *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms* (1993), 371–381.

[3] L. Holst, "On multiple covering of a circle with random arcs", *Journal of Applied Probability* **17** (1980), 284–290.

[4] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis", *Genomics* **2** (1988), 231–239.

[5] W. Feller, *An introduction to probability theory and its applications* Volume II, Second edition, John Wiley and Sons, New York, 1971.

[6] R. M. Karp, "Mapping the genome: Some combinatorial problems in molecular biology", *Proceedings of the 25th Annual ACM Symposium on Theory of Computing* (1993), 278–285.