ELSEVIER

Discrete Applied Mathematics 69 (1996) 233–245

DISCRETE APPLIED MATHEMATICS

The number of clone orderings

Lee Aaron Newberg [1]

Biological Sciences Division, The University of Chicago, 924 East 57th Street, Chicago, IL 60637-5415, USA

Received 8 July 1992; revised 14 June 1995

Abstract

This article provides an exponential generating function and a recurrence relation for computing the number of topologically distinct clone orderings.

Denote the number of maps for n clones by $c(n)$ and define the exponential generating function

$$C(x) = \sum_{n=0}^{\infty} c(n)\frac{x^n}{n!}.$$

We show $c(1) = 1$, $c(2) = 2$, $c(3) = 10$, and, for $n > 3$, that $c(n) = (4n-5)c(n-1)-(4n-7)c(n-2)+(n-2)c(n-3)$. We show

$$c(n) \sim \frac{e^{3/8}\sqrt{2}}{8n}\left(\frac{4n}{e}\right)^n.$$

We also prove that

$$C(x) = \exp\left(\frac{1+2x-\sqrt{1-4x}}{4}\right).$$

1. Introduction

With the advent of the Human Genome Project there has been much effort to construct physical and genetic maps of the chromosomes within each of us. The task is daunting and will require several years or decades of effort.

One yardstick for measuring the amount of effort required to make a physical map is the number of possible maps. This measure can take two forms. In the more abstract form, we wish to count the number of possible maps as a function of the size of the problem. That is, considering the collection of all data sets of a given size, we wish to count the number of distinguishable maps that they imply. This number is a measure of the complexity of the mapping problem. For instance, its logarithm, base

[1] Copyright 1992–1995. Research partially supported by NSF grant CCR-9017380.

SSDI 0166-218X(96)00093-3

2, is a lower bound on the worst-case running time of any algorithm that asks binary questions about a chromosome to determine the correct map.

In the second form we measure the ambiguity in a worst-case data set for a particular experimental procedure. Given a particular experiment and its data it may not be possible to reconstruct the physical map unambiguously; there may be several different maps, each consistent with the experimental data. As a function of the size of the data set this second yardstick measures the maximum over all data sets of the given size of the number of solutions consistent with that data set. This number is a measure of the quality of a particular experimental procedure in that, generally, procedures that give less ambiguous data than others are considered better.

In this article we use the first measure to gain an understanding of the Clone Ordering Problem. There are several previous results that discuss the second measure as applied to related problems. The paper [26] discusses the number of possible solutions of the Partial Digest Problem. It shows that in the worst cases the number of solutions for data from n restriction sites (and hence $N = \binom{n}{2}$ fragment lengths) is at least $\frac{1}{2}n^{0.8107}$. The paper [23] discusses the number of possible solutions of the Probed Partial Digest Problem. In the worst cases the number of solutions for data from N fragment lengths is at least $N^{1.7286}$. Also, the papers [16, 25] provide a result for the Double Digest Problem. They show that under certain statistical assumptions there can be exponentially many solutions.

Previous theoretical work on the Clone Ordering Problem can be found in [20, 2]. Algorithms for solving the Clone Ordering Problem can be found in [1, 11, 10, 14, 22].

1.1. The biology experiment

Generally, the biologist's experiment goes as follows: Take a strand of DNA and make many copies of it. Cut up these copies of the DNA in all sorts of different ways using restriction enzymes. Throw away all but those fragments (*clones*) that are near a certain easy to manage length. The biologist takes these remaining clones and runs experiments on them to determine which clones overlap which other clones and by how much. (Two clones are said to overlap if they are copied from overlapping intervals on the original DNA.) This information is used to construct a *clone ordering* (a.k.a. *contig map*), a description of the locations of the clones along the original DNA. (See [7, 4] for more information on the biological aspects of the experiment.)

Note that the term "clone ordering" is somewhat vague and is used for maps of various levels of detail. In its most detailed form it describes the exact location of each clone along the DNA. For such a map the DNA can be represented as a line segment, usually drawn horizontally, with the clones represented as equal-length subintervals. In a less detailed form a clone ordering may describe only the relative order of the left ends of the clones along the DNA, hence the name "clone ordering."

Fig. 1. The two possibilities for two clones — the two clones either overlap or don't overlap.

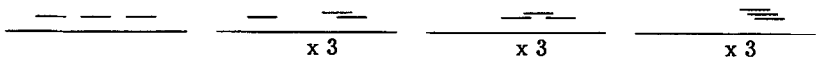

Fig. 2. The ten possibilities for three clones. The second, third, and fourth diagrams each have 3 distinguishable ways to label the clones.

There may be some portions of the DNA not represented by any of the clones. Thus, the clone ordering will show the clones in a collection of connected components known as *islands*. Those islands with more than one clone are also called *contigs*.

If we try to be too precise with our clone ordering, we may find that it is no longer a unique solution to the mapping problem. Using the experiment described there is no way to know the relative order along the DNA of the islands. Also, there is no way to tell whether a given island is present as shown or if it is present as its left–right (biologists say "sense–antisense") reflection. Furthermore, it is impossible to prove the amount of overlap of any two clones even when it is known that they do overlap.

We say that two clone orderings are *topologically similar* if one can be transformed into the other by permuting the islands and/or reflecting some of the islands. Furthermore, an adjustment of the amount by which any pair of clones overlaps leaves one with a topologically similar clone ordering if no endpoint of a clone is moved past an endpoint of another clone. We shall denote by $c(n)$ the number of topologically distinct clone orderings for n clones.

1.2. Results

For small values of n we can enumerate the maps by hand. For only one clone, there is just one map. With two clones there are two possibilities. The two clones either overlap or they do not overlap. See Fig. 1.

For three clones there are 10 possibilities. One possibility is that the three clones are mutually nonoverlapping. A second possibility is that all three clones overlap but clone #1 is between clones #2 and #3. A third possibility is that all three clones overlap but clone #2 is between clones #1 and #3, and so on. See Fig. 2 for all 10 possibilities.

The value of $c(n)$ for the first few values of n is given in Table 1. These values were computed using the values for $c(1)$, $c(2)$, and $c(3)$ and the recurrence relation we derive,

$$c(n) = (4n-5)c(n-1) - (4n-7)c(n-2) + (n-2)c(n-3), \quad n \geqslant 4.$$

We define an exponential generating function

$$C(z) = \sum_{n=0}^{\infty} c(n)\frac{z^n}{n!}$$

Table 1
The number of topologically distinct clone orderings for n distinguishable equal-length clones.

n	$c(n)$
0	1
1	1
2	2
3	10
4	94
5	1,286
6	22,876
7	499,612
8	12,925,340
9	386,356,924
10	13,099,953,016

and show that

$$C(z) = \exp\left(\frac{1 + 2z - \sqrt{1 - 4z}}{4}\right).$$

$C(z)$ gives us the asymptotic growth of $c(n)$. We show that

$$c(n) \sim \frac{e^{3/8}\sqrt{2}}{8n}\left(\frac{4n}{e}\right)^n.$$

2. Mathematical model

We model the DNA as the real line drawn horizontally. Each clone is modeled by an interval of unit length. The clones are distinguishable.

We say that two clones overlap if their intervals have a non-empty intersection. We assume that in the collection of endpoints of intervals there are no duplicates. Thus, we do not have to define whether the intervals are closed or open.

An ordered n-tuple of locations for the left endpoints of the respective intervals is called a *placement* of the clones. We define an equivalence relation on placements where the equivalence classes are called *interleavings*. Two placements are equivalent (i.e., topologically similar) if one can be transformed to the other using only the following three types of transformations:

1. *Clone sliding*: The n clones may be moved if the linear ordering of their $2n$ endpoints remains unchanged.
2. *Reflection Symmetry*: Any of the connected components (*islands*) in the set which is the union of the clones may be reflected in place.
3. *Island Re-ordering Symmetry*: The order of the islands may be permuted.

The goal is to count the interleavings.

3. Combinatorics

We shall derive the results using a three-step process. In the first step we shall assume that the clones are indistinguishable and form one island, and we shall ignore the reflection symmetry. In the second step we shall correct for distinguishability and the reflection symmetry. In the third step we shall allow more than one island.

3.1. Step 1

For this first step, assume that the clones are indistinguishable. Also for this step, ignore the reflection symmetry for islands and assume that the clones form one island.

For an interleaving, choose a placement which represents it. Label the clones 1 to n from left to right based upon the relative order of their left endpoints. Label the $2n$ clone endpoints $e_1 \leqslant \cdots \leqslant e_{2n}$ from left to right. These endpoints define $2n-1$ *atomic intervals* $(e_1,e_2),(e_2,e_3),\ldots,(e_{2n-1},e_{2n})$.

Lemma 1. *For any k, the set of clones containing the atomic interval (e_{k-1},e_k) is $\{i,i+1,\ldots,j\}$ for some $1 \leqslant i \leqslant j \leqslant n$. Furthermore, the set of clones containing (e_k,e_{k+1}) is either $\{i,\ldots,j+1\}$ or $\{i+1,\ldots,j\}$.*

Proof. Because the clones are labeled by the relative order of their left endpoints, the set of clones with left endpoints before (e_{k-1},e_k) is $\{1,\ldots,j\}$ for some $1 \leqslant j \leqslant n$. Because the clones are equilength, the order of their right endpoints is the same as the order of their left endpoints. Thus, the set of clones with right endpoints before (e_{k-1},e_k) is $\{1,\ldots,i-1\}$ for some $0 \leqslant i-1 \leqslant j$ where $i=1$ corresponds to the empty set. Because the island is connected, $i-1$ must be strictly less than j.

The endpoint e_k is either a left endpoint or a right endpoint, hence the set of clones containing (e_k,e_{k+1}) is either $\{i,\ldots,j+1\}$ or $\{i+1,\ldots,j\}$. □

We define a function P from the atomic intervals to the set $\{1,\ldots,n\}\times\{1,\ldots,n\}$ where an atomic interval gets mapped to the point (i,j) if i is the first clone containing the atomic interval and j is the last clone containing the atomic interval. The sequence $P((e_1,e_2)),\ldots,P((e_{2n-1},e_{2n}))$ which for brevity we shall denote by $P(\cdot)$, is such that each point is derived from the previous point by incrementing either the first or the second coordinate. The sequence starts at $(1,1)$ ends at (n,n) and can only include points (x,y) for which $x \leqslant y$.

Notice that this sequence does not depend on the placement chosen to represent the interleaving. Thus we have a function from interleavings to sequences. Furthermore, the function is bijective (i.e., one-to-one and onto) onto a subset of the sequences.

Lemma 2. *If $P(\cdot)$ is a sequence starting at $(1,1)$, ending at (n,n), always incrementing exactly one of the coordinates, and only passing through points (i,j) with $i \leqslant j$, then there exists a unique interleaving that gives rise to $P(\cdot)$ under the above mapping.*

Proof. Omitted. □

Thus we can count interleavings by counting these sequences. Let $a(n)$ be the number of such sequences. It is a Catalan number (see [21, 6, 13, 15, 5]) and is given by

$$a(n) = \frac{1}{n}\binom{2n-2}{n-1}.$$

It will be convenient to define a generating function (see [28, 18, 17, 12]) for $a(n)$. Let

$$A(z) = \sum_{n=1}^{\infty} a(n)z^n.$$

$A(z)$ is an analytic function well defined when $|z| < 1/4$. In this neighborhood of the origin of the complex plane we have that

$$A(z) = \sum_{n=1}^{\infty} \frac{1}{n}\binom{2n-2}{n-1} z^n = \frac{1-\sqrt{1-4z}}{2}$$

where $\sqrt{1-4z}$ is interpreted as the analytic function whose square is $1-4z$ and which equals 1 at $z = 0$. This identity is not hard to verify using repeated differentiation. Also see [19].

$A(z)$ is called a generating function for the sequence $\{a(n)\}$ because the values of the sequence can be generated from $A(z)$; in this case, the value $a(n)$ can be found through repeated differentiation of $A(z)$, which is analytic in a neighborhood of the origin. Every generating function in this article will be analytic in a neighborhood of the origin and its sequence of coefficients can be recovered through differentiation.

3.2. Step 2

Our next step is the return of distinguishability to the clones. Also, we shall now properly account for the reflection symmetry. However, we shall still require that the clones form one island.

Lemma 3. *Let $b(n)$ be the number of ways that n distinguishable equal-length clones can be interleaved to form one island. Define $B(z)$, the exponential generating function for $b(n)$, to be the exponential series*

$$B(z) = \sum_{n=0}^{\infty} b(n)\frac{z^n}{n!}.$$

Then,

$$b(n) = \begin{cases} n & \text{if } n \leqslant 1, \\ a(n)\dfrac{n!}{2} & \text{otherwise.} \end{cases}$$

Furthermore,

$$B(z) = \frac{1 + 2z - \sqrt{1 - 4z}}{4}.$$

We call $B(z)$ an *exponential* series because each $b(n)$ is multiplied by $z^n/n!$ rather than just z^n, as was the case in the definition of $A(z)$. This is why $B(z)$ simplifies to an expression quite similar to that for $A(z)$.

Proof of Lemma 3. Consider an island composed of n indistinguishable clones. It is topologically symmetric if its sequence has the property

$$(i,j) \in \{P(\cdot)\} \iff (n+1-j, n+1-i) \in \{P(\cdot)\}.$$

The clones in an asymmetric island can be labeled in $n!$ ways. If the island is symmetric and $n > 1$ then the clones can be labeled in only $n!/2$ topologically distinct ways because of the reflection symmetry. Notice that because we ignored the reflection symmetry previously, $a(n)$ double-counts asymmetric islands (of indistinguishable clones) but correctly counts symmetric ones.

Thus when $n > 1$, we have that $a(n)n!/2$ is the number of ways that n distinguishable equal-length clones can be interleaved to form one island. When $n = 1$ there is only one interleaving possible. It is impossible to form one island with no clones so $b(0) = 0$. Thus,

$$B(z) = \sum_{n=0}^{\infty} b(n)\frac{z^n}{n!} = z + \sum_{n=2}^{\infty} \frac{a(n)n!}{2}\frac{z^n}{n!} = \frac{1 + 2z - \sqrt{1 - 4z}}{4}. \quad \square$$

3.3. Step 3

We are now ready to address the "real" problem. We wish to count how many interleavings are possible when we do not restrict the number of connected components (*i.e.*, islands) that the n clones form.

Lemma 4. *Let $c(n)$ be the number of interleavings (involving any number of islands) for n clones. Define $C(z)$, the exponential generating function for $c(n)$, to be the exponential series*

$$C(z) = \sum_{n=0}^{\infty} c(n)\frac{z^n}{n!}.$$

Then,

$$C(z) = e^{B(z)} = \exp\left(\frac{1 + 2z - \sqrt{1 - 4z}}{4}\right).$$

Proof. Consider the exponential generating function $\exp(B(z))$:

$$\exp(B(z)) = \sum_{k=0}^{\infty} \frac{1}{k!} \prod_{i=1}^{k} \left(\sum_{n_k=0}^{\infty} b(n_k) \frac{z^{n_k}}{n_k!} \right)$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{n=0}^{\infty} \left\{ \sum_{\left(\sum_{i=1}^{k} n_i\right)=n} \binom{n}{n_1\ n_2\ \ldots\ n_k} \left(\prod_{i=1}^{k} b(n_i) \right) \frac{z^n}{n!} \right\}$$

$$= \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{\infty} \frac{1}{k!} \left[\sum_{\left(\sum_{i=1}^{k} n_i\right)=n} \binom{n}{n_1\ n_2\ \ldots\ n_k} \left(\prod_{i=1}^{k} b(n_i) \right) \right] \right\} \frac{z^n}{n!}$$

where

$$\binom{n}{n_1\ n_2\ \ldots\ n_k}$$

is the multinomial coefficient

$$\frac{n!}{\prod_{i=1}^{k} n_i!}$$

The coefficient of $z^n/n!$ in a single term of the innermost summation is of the form

$$\binom{n}{n_1\ n_2\ \ldots\ n_k} \left(\prod_{i=1}^{k} b(n_i) \right).$$

The multinomial coefficient counts the number of ways to allocate n distinguishable clones among k islands where the ith island from the left gets n_i clones. Conditioned upon this distribution, each $b(n_i)$ factor counts how many ways the clones allocated to the ith island can be arranged within that island. Thus, this term counts the number of ways n clones can be distributed within k ordered islands with sizes $n_1, \ldots, n_k$. The summation over all positive n_i subject to $(\sum n_i) = n$ gives a count of the number of ways n clones can be distributed within k ordered islands, regardless of their sizes. The $k!$ divisor cancels the overcounting we have introduced by ignoring the equivalence of interleavings in which the islands are permuted. (Notice that the $k!$ divisor is appropriate even if two or more of the k islands have the same size because these islands are distinguished by the clones they contain.) The summation over all k gives the number of ways n clones can be distributed within any number of unordered islands. That is, it is the number of interleavings for n clones. □

The use of exponentials in exponential generating functions is also described in [12, Example 5.5, p. 287].

4. Recurrence relation and asymptotic growth

4.1. Recurrence relation

Because of its complexity we cannot provide a closed-form expression describing each coefficient of the exponential series expansion of $C(z)$. However, we can find an easy recurrence relation by using the fact that $C(z)$ satisfies a simple differential equation.

Theorem 1.

$$c(n) = (4n-5)c(n-1) - (4n-7)c(n-2) + (n-2)c(n-3)$$

for $n \geqslant 3$.

Proof. The first two derivatives of $C(z)$ are

$$\frac{dC(z)}{dz} = \left(\frac{1}{2} + \frac{1}{2\sqrt{1-4z}}\right) C(z),$$

$$\frac{d^2C(z)}{dz^2} = \left(\frac{1}{4} + \frac{1}{2\sqrt{1-4z}} + \frac{1}{4(1-4z)} + \frac{1}{(1-4z)^{3/2}}\right) C(z).$$

We solve for $C(z)/\sqrt{1-4z}$ in the first equation and substitute into the second equation. We see that $C(z)$ satisfies the differential equation

$$(1-4z)\frac{d^2C}{dz^2} + (4z-3)\frac{dC}{dz} + (1-z)C = 0.$$

Substituting in the definition for $C(z)$ we get

$$(1-4z)\sum_{n=2}^{\infty} c(n)\frac{z^{n-2}}{(n-2)!} + (4z-3)\sum_{n=1}^{\infty} c(n)\frac{z^{n-1}}{(n-1)!} + (1-z)\sum_{n=0}^{\infty} c(n)\frac{z^n}{n!} = 0.$$

A power series can only be zero if the coefficient of every term is zero. A little algebra gives the desired result. □

This relation provides a way to compute $c(n)$ using $\Theta(n)$ arithmetic operations. The first few values of $c(n)$ can be found in Table 1. Note that this sequence is *not* in Sloane's *Handbook of Integer Sequences* [27].

4.2. Asymptotic Growth

We can calculate the asymptotic growth of $c(n)$ by examining properties of the convergence of the Taylor series of $C(z)$ about the origin. It should not be too surprising that this can be done; the reverse is done in calculus when the radius of convergence of a power series is determined through an examination of the growth of its coefficients.

Let $F(z) = \sum_{n=0}^{\infty} f(n)z^n$ be a power series. We shall say that $F(z)$ *has converging coefficient ratios* if the limit

$$f = \lim_{n\to\infty} \frac{f(n-1)}{f(n)}$$

exists and is finite but not zero. We shall call the limiting value the *limiting coefficient ratio* of $F(z)$. Note that the absolute value of the limiting coefficient ratio is the radius of convergence of $F(z)$. For example, $F(z) = \sum \pi^n 2^{\sqrt{n}} n^{\pi^2} \ln(n+\sqrt{n})z^n$ has converging coefficient ratios with a limiting coefficient ratio of $1/\pi$.

To prove the asymptotic growth of $c(n)$ we shall rely on the following two lemmas. Lemma 5 is easy to prove and we omit the details. Lemma 6 is a known result. See [24, Exercise 178], [3,9]. Also see [28, Sections 5.2 and 5.3] for similar lemmas which are applicable when $F(z)$ is restricted to other classes of functions.

Lemma 5. *If* $F(z) = \sum_{n=0}^{\infty} f(n)z^n$ *has converging coefficient ratios with limiting coefficient ratio* f, $G(z) = \sum_{n=0}^{\infty} g(n)z^n$ *has a radius of convergence* $g > |f|$, *and* $D(z) = \sum_{n=0}^{\infty} d(n)z^n$ *satisfies* $D(z) = G(z) + F(z)$ *then* $D(z)$ *has converging coefficient ratios with limiting coefficient ratio* f *and* $d(n) \sim f(n)$.

Lemma 6. *If* $F(z) = \sum_{n=0}^{\infty} f(n)z^n$ *has converging coefficient ratios with limiting coefficient ratio* f, $G(z) = \sum_{n=0}^{\infty} g(n)z^n$ *has a radius of convergence* $g > |f|$, $D(z) = \sum_{n=0}^{\infty} d(n)z^n$ *satisfies* $D(z) = G(z)F(z)$, *and* $G(f) \neq 0$ *then* $D(z)$ *has converging coefficient ratios with limiting coefficient ratio* f *and* $d(n) \sim G(f)f(n)$.

Armed with these lemmas we can prove the asymptotic growth of $c(n)$.

Theorem 2.

$$c(n) \sim \frac{e^{3/8}\sqrt{2}}{8n}\left(\frac{4n}{e}\right)^n$$

Proof. We rewrite $C(z)$:

$$\begin{aligned} C(z) &= \exp\left(\frac{1+2z}{4}\right)\left[\cosh\left(\frac{\sqrt{1-4z}}{4}\right) - \sinh\left(\frac{\sqrt{1-4z}}{4}\right)\right] \\ &= \exp\left(\frac{1+2z}{4}\right)\cosh\left(\frac{\sqrt{1-4z}}{4}\right) - \exp\left(\frac{1+2z}{4}\right)H\left(\frac{\sqrt{1-4z}}{4}\right)\frac{\sqrt{1-4z}}{4} \end{aligned}$$

where $H(y)$ is defined to be $\sinh(y)/y$ for $y \neq 0$ and $H(0) = 1$. Both $\cosh(\cdot)$ and $H(\cdot)$ are even functions that "cancel out" the square-roots in their argument and they are entire functions of z (i.e., analytic over the entire complex plane). The function $\exp((1+2z)/4)$ is also entire. Thus we can apply Lemmas 5 and 6 to relate the coefficients $c(n)/n!$ to the coefficients of $F(z) = -\sqrt{1-4z}/4$ if the relatively simple function $F(z)$ has converging coefficient ratios.

To show that $F(z)$ has converging coefficient ratios, we observe that $F(z) = \frac{1}{2}A(z) - \frac{1}{4}$ (see Section 3.1) and thus

$$\lim_{n\to\infty} \frac{f(n-1)}{f(n)} = \lim_{n\to\infty} \frac{a(n-1)}{a(n)} = \lim_{n\to\infty} \frac{n(n-1)}{(2n-2)(2n-3)} = \frac{1}{4}.$$

Therefore, $F(z)$ has converging coefficient ratios and its limiting coefficient ratio is $\frac{1}{4}$.

Lemma 5 tells us that we may ignore the term $\exp((1+2z)/4)\cosh\left(\sqrt{1-4z}/4\right)$ because it is entire. We define $G(z) = \exp((1+2z)/4)H\left(\sqrt{1-4z}/4\right)$ and use Lemma 6 on the coefficients $c(n)/n!$ and the equation $C(z) = F(z)G(z)$ to compute

$$c(n) \sim n!f(n)G(1/4) \sim n!\frac{a(n)}{2}e^{3/8} \sim \frac{e^{3/8}}{2}\frac{n}{(2n)(2n-1)}\frac{(2n)!}{n!} \sim \frac{e^{3/8}\sqrt{2}}{8n}\left(\frac{4n}{e}\right)^n$$

using Stirling's formula, $n! \sim \sqrt{2\pi n}n^n e^{-n}$. □

It is interesting to note that $c(n) = e^{3/8}b(n)$ and thus, asymptotically, 69% of all maps have a single contig.

5. Open problems

We see three ways in which these results could be expanded. The third merits the most attention.

1. Remove the assumption that no endpoint of a clone coincides with the endpoint of another clone: Because DNA is made of discrete base pairs it is cut at discrete places. Furthermore, the process of cutting the DNA into pieces is not strictly a random one. (See [8].) Thus, although the probability that two endpoints coincide may be small, it cannot be assumed to be zero.
2. Remove the assumption that the clones are of equal length: Although the clones are generally of similar lengths they need not be exactly equal. The possibility that one clone may completely contain another allows clone orderings which we have not counted here.
3. Remove the assumption that the DNA has an arbitrarily large length: More often than not the experiment produces so many clones that the sum of their lengths exceeds the length of the DNA. This necessarily precludes some of the interleavings we count. (Note that this effect is mitigated somewhat when not all overlaps are detectable. In many experiments an overlap between two clones is not detectable unless it is at least θ (usually 40 – 70%) of a clone's length. This reduces the *effective length* of the clones to a $(1-\theta)$ fraction of their original length. (See [20].) Thus, in the case of a DNA that is small, the number of maps that include only detectable overlaps may be larger than the number of maps that show all overlaps.)

Acknowledgements

Many thanks to Terry Speed, Farid Alizadeh, Dick Karp, Diane Hernek, David Blackston, Heidi Newberg, and the anonymous referees for questions and suggestions.

References

[1] F. Alizadeh, R.M. Karp, L.A. Newberg, and D.K. Weisser, Physical mapping of chromosomes: A combinatorial problem in molecular biology, Algorithmica, 13(1–2): (January–February 1995) 52–76.
[2] R. Arratia, E.S. Lander, S.Tavare and M.S Waterman, Genomic mapping by anchoring random clones — a mathematical analysis, Genomics 11(4): (December 1991) 806–827.
[3] E.A. Bender, Asymptotic methods in enumeration, SIAM Rev., 16(4): (October 1974) 485–515.
[4] E. Branscomb, T. Slezak, R. Pae, D. Galas, A.V. Carrano and M. Waterman, Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries, Genomics 8(2): (October 1990) 351–366.
[5] W.G. Brown, Historical note on a recurrent combinatorial problem, Amer. Math. Monthly. 72(9): (November 1965) 973–977.
[6] D.M. Campbell, The computation of Catalan numbers, Math. Mag., 57(4): (September 1984) 195–208.
[7] A.V. Carrano, P.J. de Jong, E. Branscomb, T. Slezak and B. Watkins, Constructing chromosome- and region-specific cosmid maps of the human genome, Genome 31(2): (1989) 1059–1065.
[8] G.A. Churchill, D.L. Daniels, and M.S Waterman, The distribution of restriction enzyme sites in Escherichia-Coli, Nucleic Acids Res., 18(3): (11 February 1990) 589–597.
[9] K.J. Compton, Some methods for computing component distribution probabilities in relational structures, Discrete Math. 66(1–2): (August 1987) 59–77.
[10] A.G. Craig, D. Nizetic, J.D. Hoheisel, G. Zehetner and H. Lehrach, Ordering of cosmid clones covering the Herpes simplex virus type-I (HSV-I) genome — a test case for fingerprinting by hybridisation, Nucleic Acids Res. 18(9): (11 May 1990) 2653–2660.
[11] A.J. Cuticchia, J. Arnold and W.E. Timberlake, ODS: Ordering DNA sequences — a physical mapping algorithm based on simulated annealing, Comput. Appl. Biosci., 9(2): (April 1993) 215–219.
[12] P. Doubilet, G.-C. Rota and R. Stanley, On the foundations of combinatorial theory (vi): The idea of generating function, Proc. of the 6th Berkeley Symposium on Mathematical Statistics and Probability 6(2): 1972 267–318.
[13] R.B. Eggleton and R.K. Guy. Catalan strikes again! How likely is a function to be convex? Math. Mag. 61(4): (October 1988) 211–219.
[14] G.A. Evans and K.A. Lewis, Physical mapping of complex genomes by cosmid multiplex analysis, Proc. Nat. Acad. Sci., USA, 86(13): (July 1989) 5030–5034.
[15] Martin Gardner, Mathematical games, Sci. Amer., 234(6): (June 1976) 120–125.
[16] L. Goldstein and M.S. Waterman, Mapping DNA by stochastic relaxation, Adv. Appl. Math., 8: (1987) 194–207.
[17] R.L. Graham, D.E. Knuth and O. Patashnik, Concrete Mathematics: A Foundation for Computer Science (Addison-Wesley, Reading, MA. 1989).
[18] D.H. Greene and D.E. Knuth, Mathematics for the Analysis of Algorithms, Progress in Computer Science and Applied Logic, Vol. 1 (Birkhauser, Boston, 3rd edition, 1990).
[19] D. A Klarner, Correspondence between plane trees and binary sequences, J. Combin. Theory 9(4): (December 1970) 401–411.
[20] E.S. Lander and M.S. Waterman, Genomic mapping by fingerprinting random clones: A mathematical analysis, Genomics, 2(3): (April 1988) 231–239.
[21] L. Lovasz, Combinatorial Problems and Exercises. (North-Holland, New York, 1979).
[22] L.A. Newberg, Finding, Evaluating, and Counting DNA Physical Maps, Ph.D. Thesis, University of California, Berkeley, CA 94720, (December 1993).
[23] L.A. Newberg and D. Naor, A lower bound on the number of solutions to the probed partial digest problem, Adv. Appl. Math. 14(2): (June 1993) 172–183.
[24] G. Pólya and G. Szegö, Problems and Theorems in Analysis, Vol. 1. (Springer, New York, 1972). Translation by D. Aeppli.

[25] W. Schmitt and M.S. Waterman, Multiple solutions of DNA restriction mapping problems, Adv. Appl. Math. 12(4): (December 1991) 412–427.
[26] S.S. Skiena, W.D. Smith and P. Lemke, Reconstructing sets from interpoint distances, in: Proceedings of the 6th Annual Symposium on Computational Geometry, Berkeley, CA, 6–8 June 1990 (ACM Press, New York, 1990) 332–339.
[27] N.J. Alexander Sloane, A Handbook of Integer Sequences. (Academic Press, New York, 1973).
[28] H.S. Wilf, Generatingfunctionology. (Academic Press, San Diego, 2nd edition, 1994).