ADVANCES IN APPLIED MATHEMATICS 12, 412–427 (1991)

Multiple Solutions of DNA Restriction Mapping Problems*

WILLIAM SCHMITT

Department of Mathematics, Memphis State University, Memphis, Tennessee 38152

AND

MICHAEL S. WATERMAN

Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113

The construction of a restriction map of a DNA molecule from fragment length data is known to be NP hard. However, it is also known that under a simple model of randomness the number of solutions to the mapping problem increases exponentially with the length of the DNA molecule. In this paper we define a hierarchy of equivalence relations on the set of all solutions to the mapping problem and study the combinatorics and characterization of the equivalence classes.

1. INTRODUCTION

Restriction maps are one of the most fundamental data structures in molecular biology. These maps show the order and location of sites (small, specific sequences) at which restriction enzymes cut double stranded molecules of DNA. For example, the restriction enzyme HhaI cuts at the sequence *GCGC*. Almost all of the several hundred known restriction enzymes cut at sequences of length 4, 6, or 8. These enzymes were only discovered in 1970, and they have given biologists a powerful tool with which to organize, manipulate, and analyze DNA.

As soon as a DNA sequence (a finite word over a four letter alphabet $\{A, C, G, T\}$) is known, restriction maps easily can be produced by inspection. Even so, before computers were so widely used, biologists occasionally overlooked an enzyme site with unfortunate consequences to subsequent experiments. There are of course many programs to convert a DNA sequence into a restriction map. However, restriction maps are often constructed before a DNA sequence is determined. These maps are sometimes preparatory in determining the sequence of the DNA, but their construction also might be the first step in other experiments. See [6] for a review.

*This research was supported by grants from the National Institutes of Health and the National Science Foundation.

0196-8858/91 $7.50

Many biologists are presently involved in genomic analysis. A genome refers to all the DNA of an organism. Until recently small segments of length 100 to 10,000 letters were most often analyzed. To organize genomic DNA, one approach is to make restriction maps of manageably small pieces and to utilize these maps to determine overlaps of pieces and thus construct a map that encompasses large parts of the genome. Kohara *et al.* [4] have successfully used this strategy to map the entire genome of *E. Coli*. Lander and Waterman [5] present a mathematical analysis of this process, and one of their conclusions is that maps should be as detailed and of as long a region as possible.

Interesting and difficult mathematical questions arise in connection with the construction of restriction maps. There are several experimental approaches to restriction mapping, each with its own advantages and disadvantages. Here we will concern ourselves the problem of mapping positions of the sites of two restriction enzymes. One way such a map is constructed in practice is by measuring the fragment lengths (not order) from a digestion of the DNA by each of the two enzymes singly, and then by two the enzymes applied together. The problem of determining the positions of the cuts from fragment length data is known as the double digest problem (DDP). In Fitch *et al.* [1] the map construction problem is approached via the set partition problem: how to choose subsets of the double digest fragments whose lengths consistently add up to the single digest fragment lengths. In Goldstein and Waterman [3] the problem is approached by a heuristic for the traveling salesman problem, stochastic annealing.

How hard is DDP restriction mapping? One answer is given by Goldstein and Waterman [3] who prove that it is NP hard. Therefore a heuristic must be used. While approximate solutions might seem easily obtainable, as in many variants of the traveling salesman problem, the situation here is more problematic. A molecular biologist wishes to find *the* correct map, the map consistent with the unknown DNA sequence. Therefore a map that is "close" to optimal as measured by some arbitrary objective function might be very far from acceptable to a biologist. Mapping algorithms should produce the smallest possible set of maps that reliably include the biologically correct map.

In addition to showing the DDP is NP hard, Goldstein and Waterman prove another disturbing result. When the enzyme sites are modelled by a random process, the number of solutions (orderings of the single digest fragments) that produce the same double digest fragments is shown to increase exponentially as the length of the DNA increases. Thus, not only is it NP hard to find an answer, but there are an exponential number of mathematically correct answers, only one of which is biologically correct. The results described here hold for exact measurements of DNA length; the large measurement errors in these data which occur in practice only compound the difficulties we have pointed out.

The object of this paper is not to produce a better algorithm for mapping DNA but to look more closely at the multiplicity of solutions. The proof in [3] depends on the Kingman subadditive ergodic theorem and is not constructive. Therefore nothing is known about the classification and combinatorics of multiple solutions.

We begin by setting up a mathematical framework for the double digest problem. The ordered set of integers from 1 to n is our model for a piece of DNA, and a (single enzyme) restriction map is a partition of this set into intervals. The set of all such partitions has a natural partial ordering, which is extremely useful in that it allows us to present a clear formulation of the double digest problem using the language of partially ordered sets. We then define a hierarchy of equivalence relations on the set of all solutions to a double digest problem, each of which partitions the set into classes of solutions which are indistinguishable from each other, given that one has a particular amount of experimental data about the problem. In some cases it is very easy to describe, and count, the members of these classes. But for one of these equivalence relations, the classes of indistinguishable solutions become very complicated, actually NP-hard to specify. We give a partial description of these classes and leave it as an open problem to completely characterize their elements.

Very little mathematical work has been done on restriction maps thus far. We feel that such important data structures deserve much more attention.

2. Notation and Basic Definitions

Let i and j be positive integers. The *interval* $[i, j]$ is the set of integers k such that $i \le k \le j$. Fix an integer $n \ge 1$, and let $\mathbf{n}$ denote the ordered set $\{1, 2, \ldots, n\}$. An *ordered partition* of $\mathbf{n}$ is an ordered set $A = \{A_1, A_2, \ldots, A_k\}$ of non-empty, disjoint intervals, called *blocks* of A, whose union is $\mathbf{n}$; and where $i < j$ if and only if $x < y$ for all $x \in A_i$ and $y \in A_j$. The number of blocks of A is denoted by $|A|$, and the *type* of A is the vector $\|A\| = (a_1, a_2, \ldots)$, where a_i is the number of blocks of A having exactly i elements. The *size* $|\mathbf{a}|$ of a vector $\mathbf{a} = (a_1, a_2, \ldots)$ having finitely many nonzero a_i, is the sum Σa_i, which is equal to $|A|$ whenever $\mathbf{a}$ is the type of an ordered partition A. For convenience we will sometimes use the symbol $1^{a_1}2^{a_2}3^{a_3}\ldots$ to denote a type vector $\mathbf{a} = (a_1, a_2, \ldots)$.

Let Ω_n denote the set of all ordered partitions of $\mathbf{n}$. Ω_n is partially ordered by the relation $A \le B$ in Ω_n if and only if every block of A is contained in some block of B. A good way of understanding the partially ordered set Ω_n is by identifying each ordered partition A with the set of locations, or *cut-sites*, at which the string $\{1, 2, \ldots, n\}$ is divided in order to obtain A. We will adopt the convention of identifying a cut-site, which occurs between consecutive integers, with the integer immediately to its left. More formally, let $\mathscr{B}_{n-1}$ be the boolean algebra of all subsets of the set $\{1, 2, \ldots, n-1\}$. Construct a map $\phi\colon \Omega_n \to \mathscr{B}_{n-1}$ by letting $\phi(A) = \{\max\{A_i\} | 1 \le i \le k-1\}$, for all $A = \{A_1, A_2, \ldots, A_k\}$ in Ω_n.

PROPOSITION 1. *The mapping $\phi\colon \Omega_n \to \mathscr{B}_{n-1}$ is an anti-isomorphism of partially ordered sets.*

Proof. To see that ϕ is a bijection, we exhibit a two-sided inverse for ϕ. Let $U = \{i_1, i_2, \ldots, i_k\}$ be an element of $\mathscr{B}_{n-1}$, with $i_1 < i_2 < \cdots < i_k$. Letting $i_0 = 0$ and $i_{k+1} = n$, define $\pi(U)$ to be the ordered partition $\{A_1, A_2, \ldots, A_{k+1}\}$, where $A_j = [i_j + 1, i_{j+1}]$, for $0 \le j \le k$. It is easy to check that π is the required inverse map. Also, it is immediate from the definition of ϕ and the ordering of Ω_n that $A \le B$ in Ω_n if and only if $\phi(A) \supseteq \phi(B)$ in $\mathscr{B}_{n-1}$. Therefore ϕ is an anti-isomorphism. □

Hence Ω_n is a boolean algebra, and thus for all $A, B \in \Omega_n$, the greatest lower bound, or *meet*, of A and B exists, and the least upper bound, or *join*, of A and B exists. These operations can be expressed via the correspondence ϕ as $A \wedge B = \phi^{-1}(\phi(A) \cup \phi(B))$, and $A \vee B = \phi^{-1}(\phi(A) \cap \phi(B))$, respectively. The meet of A and B can be written explicitly as

$$A \wedge B = \{A_i \cap B_j | A_i \in A,\ B_j \in B \text{ and } A_i \cap B_j \neq \emptyset\}.$$

However, the join of A and B is most easily described as above, i.e., as the ordered partition whose set of cut-sites is the intersection of the sets of cut-sites, which we call the set of *coincident cut-sites* of the pair (A, B). The restrictions of the pair (A, B) to each of the blocks of $A \vee B$ are the *connected components*, or simply, the *components* of the pair (A, B). The number of components of (A, B) is given, via proposition 1, by

$$|A \vee B| = |A| + |B| - |A \wedge B|. \tag{1}$$

In the case when the pair (A, B) has no coincident cut-sites, the number

of blocks of the meet $A \wedge B$ is given by the formula

$$|A \wedge B| = |A| + |B| - 1.$$

3. The Double Digest Problem and Equivalence Classes of Solutions

3.1. *The Double Digest Problem*

A *restriction map* (of a single restriction enzyme) is simply a linearly ordered partition of $\mathbf{n}$. Let Ω_n^k denote the subset of Ω_n consisting of restriction maps having k blocks, or *fragments*. Given $A \in \Omega_n^k$, and a permutation $\sigma \in S_k$, a new restriction map $A^\sigma = \{A_1^\sigma, A_2^\sigma, \ldots, A_k^\sigma\}$ is uniquely defined by the condition $|A_i^\sigma| = |A_{\sigma(i)}|$ for $1 \le i \le k$. The rule $\sigma\colon A \to A^\sigma$ thus defines an action of symmetric group S_k on the set Ω_n^k. The orbit of a restriction map A under this action is the set of all restriction maps A' with $\|A'\| = \|A\|$. For the purposes of visualization, it is helpful to think of a permutation σ as actually permuting the fragments of the restriction map A in order to obtain the map A^σ. One must bear in mind, however, that σ is really only permuting the *locations* of the fragments of A, while the order of the underlying set, $\mathbf{n}$, remains fixed.

In the double digest problem, one is given a set of data consisting of a triple of vectors $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, where $\mathbf{a} = \|A\|$, $\mathbf{b} = \|B\|$, and $\mathbf{c} = \|A \wedge B\|$, for some specific (but unknown) pair of restriction maps (A, B), corresponding to a map of two restriction enzymes. The problem is then to try to recover the pair (A, B) from the given data $(\mathbf{a}, \mathbf{b}, \mathbf{c})$. As mentioned in the Introduction, there are two major difficulties that one encounters in trying to solve this problem. First of all, the solution to a particular double digest problem is usually far from unique. In fact, Goldstein and Waterman [3] showed that under a certain probability model, there are an exponentially increasing number of solutions as a function of segment length n with probability one. The second difficulty (also shown in [3]) is that the problem of finding even a single solution is NP complete.

In the following sections we study the set $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of all solutions to the double digest problem DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$ defined by some fixed set of data $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, where $|\mathbf{a}| = k$, $|\mathbf{b}| = m$, and $|\mathbf{c}| = r$. Given any pair of solutions (A, B) and (A', B') in $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$, we can always write $(A', B') = (A^\sigma, B^\pi)$, for some (not necessarily unique) pair of permutations $\sigma \in S_k$, and $\pi \in S_m$, where necessarily, $A^\sigma \wedge B^\pi = (A \wedge B)^\gamma$ for some γ in S_r. Relationships between the solutions (A, B) and (A', B') often can be expressed most easily in terms of properties of the permutations σ, π, and γ. There are several natural equivalence relations that can be defined on this set, each of which partitions $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ into classes consisting of solutions

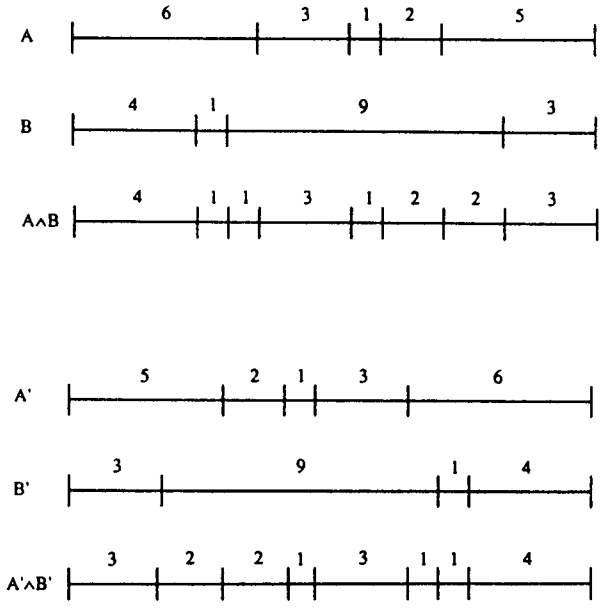

FIG. 1. Reflection of solutions.

which are indistinguishable from each other, assuming a certain level of knowledge about the problem at hand.

3.2. *Reflections and Physical Solutions*

Let $(A, B) \in \mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ be a solution, and suppose the permutations σ and π reverse the order of the sets $\{1, 2, \ldots, k\}$ and $\{1, 2, \ldots, m\}$, respectively. The pair (A^σ, B^π), called the *reflection* of (A, B), is clearly also in $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ (Fig. 1).

The pairs of restriction maps (A, B) and (A', B') in Fig. 1 are reflections of each other, and they are both solutions for the same double digest problem. In a very strong sense, they represent the same solution to the problem, since they differ only by an arbitrary choice of orientation, and no experimental data could possibly serve to distinguish one from the other. Therefore we define the set $\mathscr{S}_1(\mathbf{a}, \mathbf{b}, \mathbf{c})$ of *physical solutions* to DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$ to be the set of all solutions $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ modulo the reflection relation.

3.3. *Overlap Equivalence of Solutions*

Throughout this section, we suppose the problem DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is defined by the (unknown) pair of restriction maps (A, B), where the sizes of all the

blocks of A are distinct and the sizes of all the blocks of B are distinct. Then in addition to the data $\mathbf{a} = \|A\|$, $\mathbf{b} = \|B\|$, and $\mathbf{c} = \|A \wedge B\|$, it is possible (in principle) to obtain a complete set of *overlap data* for A and B. That is, for each block A_i of A and each B_j in B, one can determine whether or not A_i and B_j have non-empty intersection. Experimentally this can be accomplished by first digesting with enzyme A, then digesting each fragment with enzyme B, and then repeating this procedure with the roles of A and B reversed. In this way one can identify which blocks of A and B contain each block if $A \wedge B$, and from this point the overlap data can be determined easily (see [7] for details about this last step). The overlap data of a pair (A, B) is nicely represented by the *interval graph* $G(A, B)$, whose vertices are the (labelled) set of blocks $A \cup B$ and whose edges are the set of all pairs $\{A_i, B_j\}$, such that $A_i \cap B_j$ is non-empty. Interval graphs of restriction maps are studied in [7], where also a linear time algorithm is presented for finding a pair of restriction maps having a given set of overlap data. Knowing the overlap data is usually not sufficient to determine the physical solution corresponding to (A, B) uniquely; but it is a relatively simple matter to describe the classes of solutions in $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ which have the same overlap data. These will be called the classes of *overlap equivalent* solutions to DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$.

If the pair (A, B) has t connected components, then the components may be permuted in any of $t!$ ways, and any subset of the set of components may be reflected, and we will obtain a solution which is clearly overlap equivalent to (A, B). Reflecting any component which consists of one block from each of A and B does not give a new solution. Thus if there are r such components, then (A, B) is one of $2^{t-r}t!$ overlap equivalent elements of $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$ which can be obtained by such rearrangement of the components of (A, B). Note that this number is independent of the choice of (A, B) in $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$, because by Eq. (1), the number of components $t = |A \vee B|$ is the same for all (A, B) in $\mathscr{S}(\mathbf{a}, \mathbf{b}, \mathbf{c})$.

Another manner in which overlap equivalent solutions can occur is described as follows: For each $B_i \in B$, let $\mathscr{A}_i$ be the set of all integers u such that $A_u \subseteq B_i$, and for each $A_j \in A$, let $\mathscr{B}_j$ be the set of all u such that $B_u \subseteq A_j$. Notice that each of the sets $\mathscr{A}_i$ and $\mathscr{B}_j$ consists of consecutive integers and is thus an interval in the ordered set A or B. We call the $\mathscr{A}_i$ and $\mathscr{B}_j$ the *intervals of uncut fragments* of A and B, respectively. If $\sigma \in S_k$ maps each $\mathscr{A}_i$ to itself, while fixing all elements of $\{1, 2, \ldots, k\}$ which are not contained in any $\mathscr{A}_i$, and similarly, $\pi \in S_m$ permutes each $\mathscr{B}_j$ while fixing the rest of $\{1, 2, \ldots, m\}$, then the pair (A^σ, B^π) is clearly a solution which is overlap equivalent to (A, B). Any solution which is overlap equivalent to (A, B) must be obtainable by such permutations within intervals of uncut fragments and/or rearrangement of components of (A, B). The reflection of any component which contains

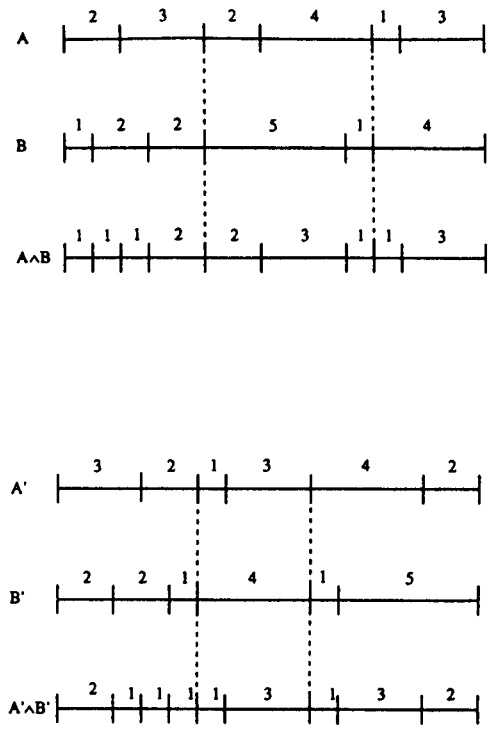

FIG. 2. Rearrangement of components.

only one block of A and/or B can be affected by a permutation of uncut fragments. Thus if (A, B) has exactly s such components then the overlap equivalence class which contains (A, B) consists of exactly

$$2^{t-s}t! \prod_{i=1}^{k} |\mathscr{A}_i|! \prod_{j=1}^{m} |\mathscr{B}_j|!$$

different solutions to DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$, where t is the number of connected components of (A, B). For example, the pairs (A, B) and (A', B') shown in Fig. 2 are two of the $2^{3-1}3!2! = 48$ different overlap equivalent solutions (all rearrangements of components) to DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$, where $\mathbf{a} = 1^1 2^2 3^2 4^1$, $\mathbf{b} = 1^2 2^2 4^1 5^1$, and $\mathbf{c} = 1^5 2^2 3^2$.

The pairs of restriction maps shown in Fig. 3 are two of $2 \cdot 2!3! = 24$ different overlap equivalent solutions (permutations within intervals of uncut fragments and reflections) to DDP$(\mathbf{a}, \mathbf{b}, \mathbf{c})$, where $\mathbf{a} = 1^1 2^1 3^1 5^1 6^1$, $\mathbf{b} = 1^1 3^1 4^1 9^1$, and $\mathbf{c} = 1^3 2^2 3^2 4^1$.

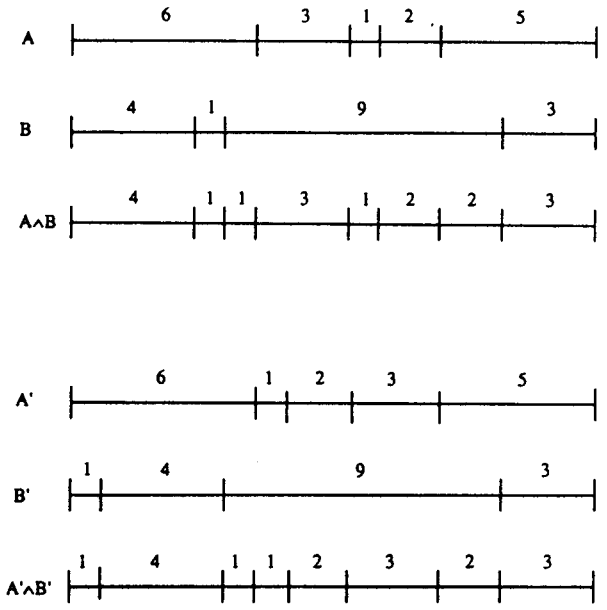

FIG. 3. Permutation within intervals of uncut fragments.

3.4. *Overlap Size Equivalence of Solutions*

When either (or both) of the restriction maps A and B contain more than one piece of the same length, the equivalence classes of potentially distinguishable solutions become much more complicated. Given a pair of restriction maps (A, B), and a block $C_s \in A \wedge B$, we have $C_s = A_{i_s} \cap B_{j_s}$ for some unique $A_{i_s} \in A$ and $B_{j_s} \in B$. The *overlap size data* of the pair (A, B) is defined to be the (unordered) set of ordered triples of integers

$$\left\{\left(|C_s|, |A_{i_s}|, |B_{j_s}|\right) \middle| C_s \in A \wedge B\right\}.$$

Two solutions to DDP(**a**, **b**, **c**) are said to be *overlap size equivalent* if they have the same set of overlap size data. In the case where each of A and B consist of fragments of distinct lengths, knowing the overlap size data is equivalent to knowing the complete set of overlap data for (A, B). But when A or B contains multiple fragments of the same length, the overlap size data gives less information about the pair (A, B) than the complete set of overlap data and is all that can be determined from the experiments described above. This loss of map information corresponds to our inability, using such experiments, to separate and thus distinguish between different pieces of DNA having the same length in a given digest.

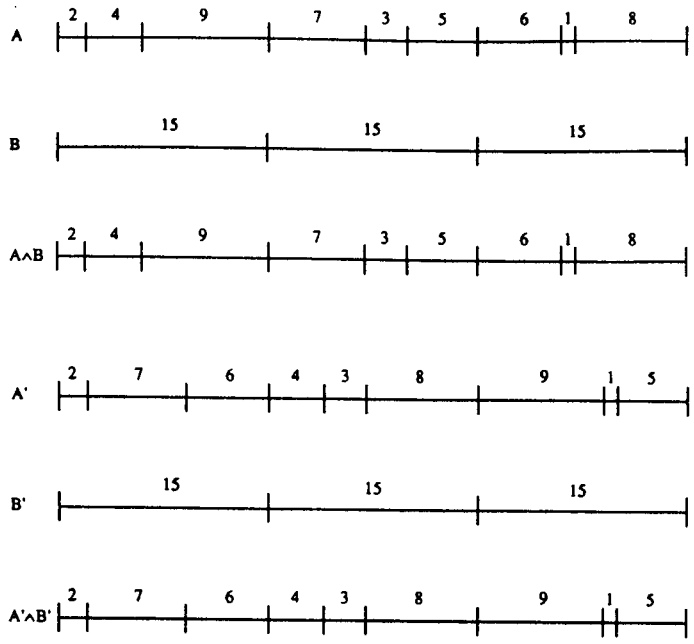

FIG. 4. Permutation of uncut fragments.

Given a solution (A, B), the problem of describing the set of all solutions which are overlap size equivalent to (A, B) is much more difficult than describing those solutions which are overlap equivalent to (A, B). For example, in Fig. 4 the overlap equivalence classes of the pairs (A, B) and (A', B') are disjoint from each other, each containing $2^{3-3}(3!)^3 = 216$ pairs, while (A, B) and (A', B') are overlap size equivalent.

This simple example indicates one of the essential difficulties in trying to describe the overlap size equivalence class of an arbitrary pair of restriction maps (A, B): the uncut fragments no longer need be permuted only within intervals. Suppose that fragments B_i and B_j of B have the same length. Let $\mathscr{A}_i$ and $\mathscr{A}_j$ be the intervals of uncut fragments of A contained in B_i and B_j, respectively, and let L_i be the sum of the lengths of the fragments in $\mathscr{A}_i$. Then in the process of finding all solutions which are overlap size equivalent to (A, B), one must determine all subsets S of $\mathscr{A}_i \cup \mathscr{A}_j$ such that the sum of the lengths of the elements of S is equal to L_i. But this is a version of the set partition problem (see [2]), which is known to be NP-complete.

Given a pair of restriction maps (A, B) and an interval $I_C \subseteq A \wedge B$, the *cassette* defined by I_C is the pair of intervals $\mathscr{C} = (I_A, I_B)$, where I_A and I_B are the sets of all blocks of A and B, respectively, which contain a block

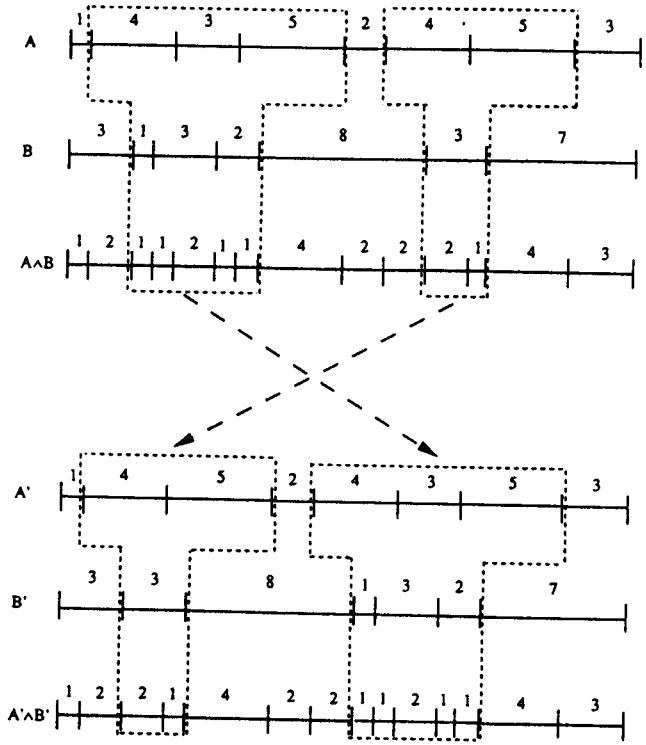

FIG. 5. Exchange of cassettes.

of I_C. Let m_a and m_b in **n** be the minimal elements of the left-most blocks of I_A and I_B, respectively. If $m_a \neq m_b$, then the *left end fragment* of $\mathscr{C}$ is the block of I_A or I_B which contains the smaller of these numbers. If $m_a = m_b$ then the cassette $\mathscr{C}$ has no left end fragment. The *left overlap* of $\mathscr{C}$ is the distance $|m_a - m_b|$. The *right end fragment* and *right overlap* of $\mathscr{C}$ are defined similarly, by substituting the words "maximal" and "right-most" for the words "minimal" and "left-most" in the above.

Suppose two cassettes $\mathscr{C}$ and $\mathscr{C}'$ within a solution (A, B) to DDP(**a**, **b**, **c**) have left end fragments and overlaps of the same length and right end fragments and overlaps of the same length. Then these cassettes may be exchanged as in Fig. 5, and one obtains a new solution (A', B') which is overlap size equivalent to (A, B).

Also, if the left and right end fragments of a single cassette $\mathscr{C}$ in (A, B) have the same length, and the left and right overlaps are the same size, then the cassette may be reversed or *reflected* as in the example of Fig. 7. The components of a pair of restriction maps are special examples of

cassettes (they are cassettes having no end fragments), and thus rearrangements of components are special cases of compositions of exchanges and reflections of cassettes.

As we have seen in the last section, the overlap equivalence classes of solutions are generated by permutations within intervals of uncut fragments and rearrangements of components. We have just described generalizations of these two types of permutations which preserve the overlap size data of a solution. An interesting question to ask at this point is: What other types (if any) of fundamental "moves" are needed in order to generate the entire overlap size equivalence class of a solution?

4. An Example

The general occurrence of multiple solutions to the double digest problem was discovered in [7] where the stochastic annealing algorithm was tested on DDP(**a**, **b**, **c**), where $\mathbf{a} = 1^1 3^2 12^1$, $\mathbf{b} = 1^1 2^1 3^2 4^1 6^1$, and $\mathbf{c} = 1^4 2^3 3^1 6^1$. The data was obtained from the pair of maps (A, B) in Fig. 6. The algorithm returned the pair of maps (A', B') from Fig. 6, which is also a solution to DDP(**a**, **b**, **c**). In this section we examine this example in more detail.

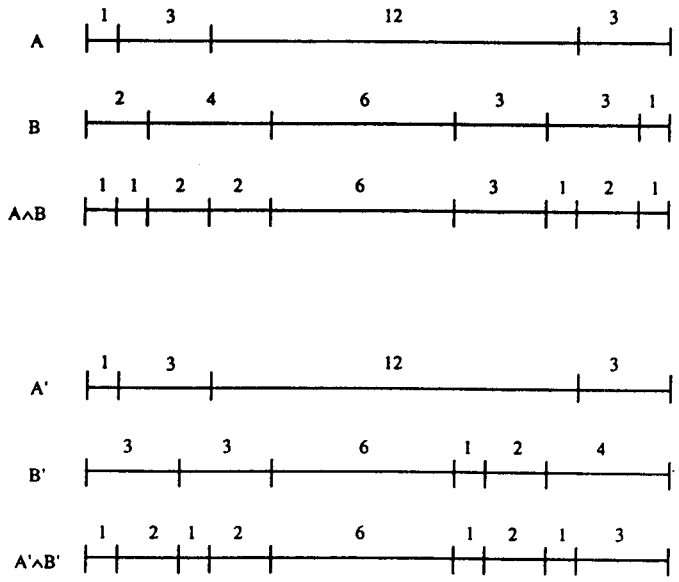

FIG. 6. An example of multiple solutions from [3].

This problem has 208 different solutions which fall into 26 different overlap equivalence classes: 13 classes with 4 members each, and 13 classes of 12 members. The solution (A, B) in Fig. 6 has an overlap equivalence class containing 4 elements, which are generated by the reflection of the whole pair, and the reversal of the uncut fragments of length 3 and 6 in B. The overlap equivalence class of the solution (A', B') contains 12 elements: $3! = 6$ permutations of uncut fragments in B multiplied by a factor of 2 for the reflection of the pair.

Somewhat surprisingly, the overlap size equivalence classes do not correspond precisely to the overlap equivalence classes in this rather small problem. There are 25 overlap size equivalence classes of solutions to this DDP(**a**, **b**, **c**): 11 classes of 4 members, 13 classes of 12 members, and 1 class having 8 members. The solution (A, B) in Fig. 6 is a member of this unique class of eight solutions, which is the union of two different 4-element overlap equivalence classes, which are related by the cassette reversal illustrated in Fig. 7.

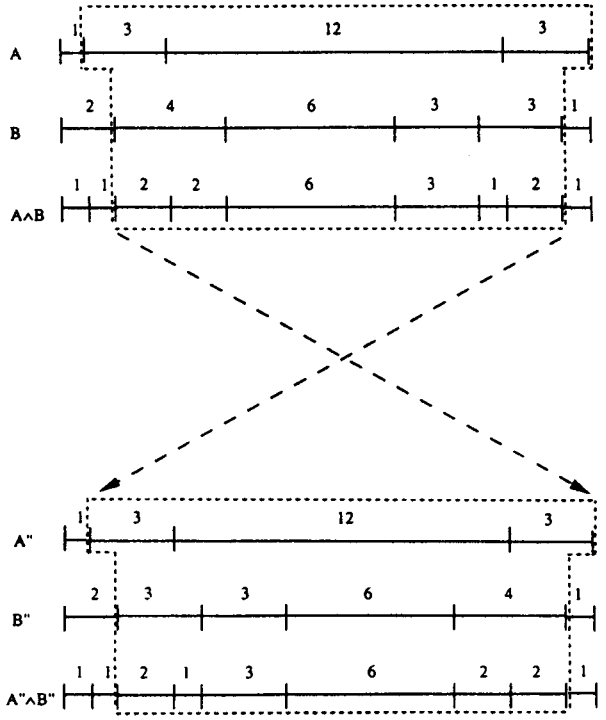

FIG. 7. Cassette reversal.

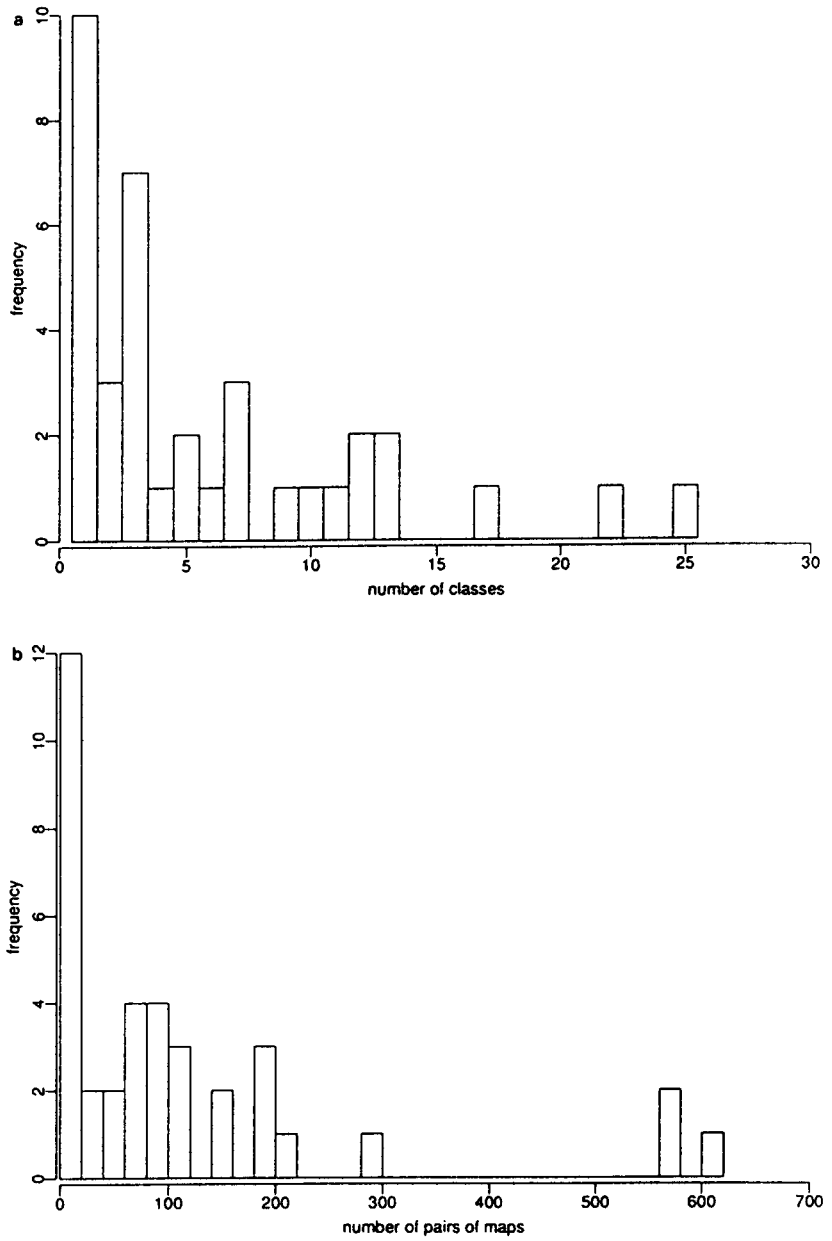

FIG. 8. (a) Frequency of occurrence of problems with a given number of overlap size equivalence classes of solutions. (b) Frequency of occurrence of problems with a given number of solutions.

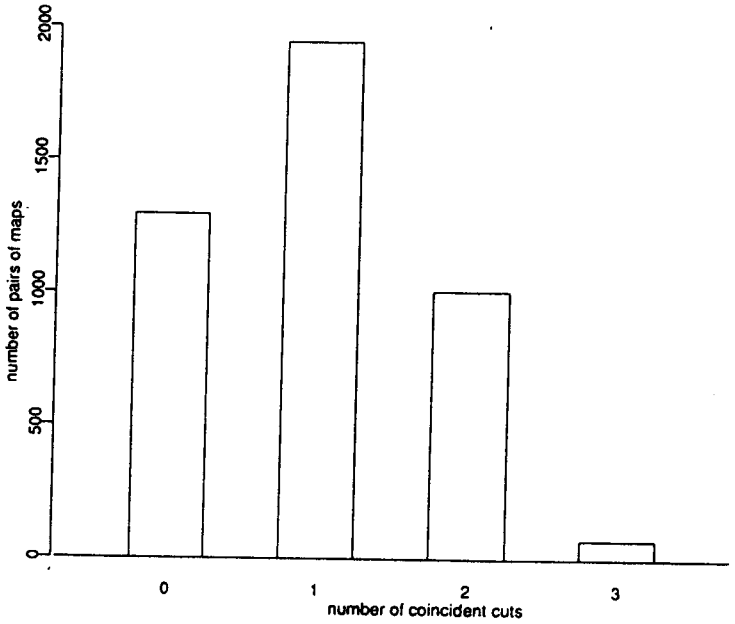

FIG. 9. Number of pairs of maps with a given number of coincident cuts.

We now consider the set of all pairs of restriction maps (A, B) with $\|A\| = \mathbf{a} = 1^1 3^2 12^1$ and $\|B\| = \mathbf{b} = 1^1 2^1 3^2 4^1 6^1$. There are $4!6!/2!2! = 4320$ different pairs having this data, with 37 different vectors $\mathbf{c} = \|A \wedge B\|$; i.e., there are 37 different double digest problems DDP($\mathbf{a}, \mathbf{b}, \mathbf{c}$) having these values of $\mathbf{a}$ and $\mathbf{b}$. Figure 8a shows the frequency of occurrence of problems in this set having a given number of overlap size equivalence classes of solutions. For example, 10 of these problems have a unique (overlap size) equivalence class of solutions, while 20 have three or fewer classes of solutions.

Figure 8b gives the frequency of occurrence of problems having a given number of pairs of restriction maps as solutions. For example, four of these problems have greater than or equal to 80 and less than 100 solutions.

It is interesting to note that the problem considered above, with $\mathbf{c} = 1^4 2^3 3^1 6^1$ is the only one of these problems having the maximal number, 25, of classes of solutions, while it is composed of 208 solutions.

Given the above values of $\|A\|$ and $\|B\|$, it is possible for the pair (A, B) to have 0, 1, 2, or 3 coincident cuts. Figure 9 shows the number of such pairs of maps having a given number of coincident cut-sites.

REFERENCES

1. W. M. FITCH, T. F. SMITH, AND W. W. RALPH, Mapping the order of DNA restriction fragments, *Gene* **22** (1983), 19–29.
2. M. R. GAREY AND D. S. JOHNSON, "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freeman, San Francisco, 1979.
3. L. GOLDSTEIN AND M. S. WATERMAN, Mapping DNA by stochastic relaxation, *Adv. Appl. Math.* **8** (1987), 194–207.
4. Y. KOHARA, A. AKIYAMA, AND K. ISONO, The physical map of the whole *E. Coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library, *Cell* **50** (1987), 495–508.
5. E. S. LANDER AND M. S. WATERMAN, Genomic mapping by fingerprinting random clones: A mathematical analysis, *Genomics* **2** (1988), 231–239.
6. D. NATHANS AND H. O. SMITH, Restriction endonucleases in the analysis and restructuring of DNA molecules, *Annual Rev. Biochem.* **44** (1975), 273–293.
7. M. S. WATERMAN AND J. R. GRIGGS, Interval graphs and maps of DNA, *Bull. Math. Biol.* **48**, No. 2 (1986), 189–195.