ADVANCES IN APPLIED MATHEMATICS 8, 194–207 (1987)

# Mapping DNA by Stochastic Relaxation*,[1]

LARRY GOLDSTEIN[†] AND MICHAEL S. WATERMAN[‡,†]

[†]*Departments of Mathematics and* [‡]*Molecular Biology, University of Southern California, Los Angeles, California 90089-1113*

The multiple digest mapping problem arising in molecular biology can be stated roughly as follows. A linear or circular segment of DNA is cut at all occurrences of a specific short pattern by restriction enzymes. By using restriction enzymes singly and in combination it is required to construct a map showing the location of cleavage sites. In this paper we first consider the efficacy of a simulated annealing algorithm towards the solution to the multiple digest problem. Second, the double digest problem, the simplest version of the multiple digest problem with only two restriction enzymes used, is shown to admit an exponentially increasing number of solutions as a function of the length of the segment under a particular probability model. Next, the double digest problem is shown to lie in the class of NP complete problems which are conjectured to have no polynomial time solution. Last, the construction of circular maps is considered and the problem of measurement error is discussed. 

## 1. INTRODUCTION

The multiple digest mapping problem arising in molecular biology can be stated roughly as follows. A linear or circular segment of DNA is cut or cleaved at all occurrences of a specific short pattern by restriction enzymes. DNA sequences can be viewed as finite sequences over the four letter alphabet $\{A, C, G, T\}$. Each restriction enzyme cuts the double stranded DNA at a short pattern specific to that enzyme; the restriction enzyme HhaI, for example, cuts at GCGC. Restriction enzymes can be used singly or in combination and the resulting fragment lengths recorded. It is required to construct a map using the fragment length data to show the location of cleavage sites.

Site specific restriction enzymes were discovered in 1970 and soon biologists were making restriction maps. (See the review [14].) These maps are fundamental to molecular biology. The first computer algorithm seems

* This work was supported by grants from the System Development Foundation and from the National Institutes of Health.

[1] Dedicated to Nick Metropolis



0196-8858/87 $7.50

to be that of Stefik [19] who applied concepts from artificial intelligence. Pearson [16] then proposed solving the two single–one double digest problem, referred to as problem DDP, by considering all permutations of the two single digests. Fitch, Smith, and Ralph [5] present an algorithm for problem DDP which is based on the additive relationship between the double and single digests. Nolan, Mairna, and Szalay [15] also approach automatic mapping via these relationships. Durand and Bregerere [4] modify Stefik's approach to a branch and bound algorithm. Wulkan and Llott [22] allow the biologist to supply information about the map and thereby reduce the number of permutations.

In [21], the graph theoretic nature of restriction maps is studied. Questions about the statistical distribution of occurrences and restrictions sites are studied in [2], [18] and [20]. The distributional questions are complicated by the fact that nonoverlapping occurrences are the feature of biological interest.

In this paper we first describe a simulated annealing solution to the multiple digest problem and present results of a study of its efficacy. Second, an analytic result of the nonuniqueness of solutions is presented; more specifically, we demonstrate that under a certain probability model there are an exponentially increasing number of solutions as a function of the length of the segment with probability one. Then we demonstrate the multiple digest problem to lie in the class of NP complete problems conjectured to have no polynomial time solution. Last, we point out that a simple modification of the algorithm handles mapping circular DNA, but that no effective algorithm has been devised for fragments with realistic measurement errors.

Currently there is a great deal of discussion about mapping and sequencing the human genome [11]. Technology has developed to the point that such a project may soon be started. The two largest genomes sequences to date are of approximate sizes $5 \times 10^4$ and $2 \times 10^5$ bases of DNA. The human genome is approximately $3 \times 10^9$ bases of DNA. While this paper is concerned with the usual, small mapping problems of molecular biology, it is relevant to large mapping projects. The suitability of annealing, the multiplicity of solutions (only one of which is correct) and the computational intractability of the multiple digest problem are all important considerations.

## 2. Simulated Annealing Solutions to the Multiple Digest Problem

We begin with a description of the simulated annealing algorithm. Let $V$ be a finite set of elements, and $f$ a function that assigns a real number to each element of $V$. Suppose we wish to find an element $v^* \in V$ that

corresponds to the global minimum value of $f$; that is, find $v^* \in V$ such that $f(v^*) = \min_{v \in V} f(v)$. For any $T > 0$, let $\pi_T$ be the Gibbs distribution over $V$ given by

$$\pi_T(v) = \exp\{-f(v)/T\}/Z,$$

where $Z$, the partition function, is chosen such that $\sum_{v \in V} \pi_T(v) = 1$. Note that for large values of $T$ the distribution tends to be uniform over $V$, while for small values of $T$ the favorable elements of $V$, that is, those elements of $V$ for which $f(v)$ is small, are weighted with large probability. Therefore, a probabilistic solution to the problem of locating an element $v \in V$ for which $f(v)$ is minimized is given by sampling from the distribution $\pi_T$ for small $T > 0$.

One way this may be achieved is to simulate a Markov chain $\{X_n\}_{n \geq 0}$ with state space $V$ that has $\pi_T$ as its stationary distribution and let it approach equilibrium. First, this requires that one determine for each $v \in V$ a set of neighbors $N_v \subset V$ where transitions from $v$ are allowed in such a way that the resulting Markov chain is irreducible. Let the collection $\{N_v\}$ also satisfy for all $v$, $v \in N_v$, $w \in V$,

$$v \in N_w \text{ if and only if } w \in N_v$$

and

$$|N_v| = |N_w|,$$

and now define transition probabilities $p_T(v, w) = P_T(X_{n+1} = v | X_n = w)$ by

$$p_T(v, w) = 0 \qquad \text{if } v \text{ is not in } N_w,$$

$$p_T(v, w) = \exp\{-(f(v) - f(w))^+/T\}/|N_w| \qquad \text{if } v \in N_w, v \neq w,$$

and $p_T(w, w)$ fixed by the requirement

$$\sum_{v \in N_w} p_T(v, w) = 1.$$

An easy calculation now shows that $\pi_T$ satisfies the balance equation

$$p_T(v, w)\pi_T(w) = p_T(w, v)\pi_T(v),$$

which is sufficient to guarantee that $\pi_T$ is the unique stationary distribution of the chain $X_n$.

In practice, as the function $f$ may be expensive computationally, the Markov chain is simulated in the following way: when at $w$ a neighbor of $w$ is selected from $N_w$ uniformly, say $v$, and $f(v)$ is computed. The move to $v$ is then accepted with probability

$$p = \exp\{-(f(v) - f(w))^+/T\}$$

and the new state of the chain is $v$, else the move is rejected and the state of the chain remains $w$.

This method was proposed by Metropolis *et al.* [13] and has the following statistical mechanical interpretation. The set $V$ can be thought of as the set of all possible configurations of some physical system; the quantity $f(v)$ is the energy of the system when in configuration $v$ with $T$ playing the role of temperature. The Gibbs distribution then gives the probability of finding the system in a particular configuration at some given temperature. At high temperature, the system can be found in any of its states with approximately equal probabilities while at low temperature it is more likely that the system will be in a low energy configuration.

Kirkpatrick *et al.* [10] introduced the idea of cooling the system in the hope that in the limit the distribution $\pi_0 = \lim_{T \downarrow 0} \pi_T$ will be obtained; $\pi_0$ is that distribution that distributes mass one uniformly over the states of minimum energy. In this way the algorithm resembles the physical process of annealing, or cooling, a physical system. As in the physical analog, the system may be cooled too rapidly and become trapped in a state corresponding to a local energy minimum; Geman and Geman [7] showed, in a simulated annealing algorithm pertaining to image reconstruction, that if at stage $n$ in the algorithm one uses the transition probabilities given above with temperature $T_n$, where $T_n \downarrow 0$ and $T_n \geqq c/\log(n)$ with $c$ a constant that depends on $f$, then the state of the Markov chain converges in distribution to $\pi_0$. See also work of Hajek [8].

The algorithm yields a general, that is, problem non-specific, way to attack many difficult combinatorial optimization problems. It should be noted that in order to implement the simulated annealing algorithm the user has control over the energy function and the neighborhood structure on $V$. The success or failure of the algorithm may depend on these choices.

Bonomi and Lutton [1] applied a version of the simulated annealing algorithm that they called the extended Metropolis method to the travelling salesman problem and for large problems reported their method to be competitive with Lin's 2-opt algorithm and the convex hull algorithm coupled to the 2-opt procedure.

In the travelling salesman problem, known to belong to the class of NP complete problems conjectured to have no polynomial time solution, one wishes to find a path, or tour, of minimal length to be taken by a salesman required to visit each of $n$ cities, labeled $1, 2, \ldots, n$, and then return home. The set $V$ in this case may be taken to be $S_n$, the set of all permutations of $\{1, 2, \ldots, n\}$ where to each permutation $\sigma \in S_n$ we identify the corresponding configuration given by the tour taken in the order dictated by $\sigma$. The energy may be taken to be the total length of the tour although we note that any monotone transformation of this quantity may also serve.

In [1], Bonomi and Lutton choose a neighborhood structure for $S_n$ motivated by Lin's 2-opt [12] deterministic algorithm for the travelling

salesman problem. If, for a given tour $\sigma$ we imagine links connecting neighboring cities in the tour, we say that the tour $\sigma$ is $k$-opt, $1 \leqq k \leqq n$, if for all tours that can be obtained from $\sigma$ by breaking at most $k$ links, the tour given by $\sigma$ is the shortest. Thus, every tour is 1-opt and only the true best tours are $n$-opt.

It is easily seen that a tour $\sigma = (i_1, i_2, \ldots, i_n)$ is 2-opt if and only if it yields the shortest tour of all tours which are elements of $N(\sigma) = \{\tau \in S_n: \tau = (i_1, i_2, \ldots, i_{j-1}, i_k, i_{k-1}, \ldots, i_{j+1}, i_j, i_{k+1}, \ldots, i_n)$ for some $1 \leqq j \leqq k \leqq n\}$. It is not hard to see that given any initial tour $\sigma_0$ and any final tour $\sigma_n = (j_1, j_2, \ldots, j_n)$ we may obtain $\sigma_n$ from $\sigma_0$ through a sequence of permutations $\sigma_1, \sigma_2, \ldots, \sigma_{n-1}$ such that $\sigma_k \in N(\sigma_{k+1})$ for $k = 0, 1, \ldots, n - 1$ as follows. Given $\sigma_k$ such that

$$\sigma_k = (j_1, j_2, \ldots, j_k, l_{k+1}, \ldots, l_m, l_{m+1}, \ldots, l_n)$$

where $j_{k+1} = l_m$, say, invert $l_{k+1}$ through $l_m$ to obtain

$$\sigma_{k+1} = (j_1, j_2, \ldots, j_k, j_{k+1}, l_{m-1}, \ldots, l_{k+1}, l_{m+1}, \ldots, l_n).$$

Thus we see that this notion of neighborhood yields an irreducible Markov chain in the algorithm described above. The three other requirements desired of a neighborhood structure listed above are satisfied trivially.

The problem we consider, the multiple digest problem, is as follows. We discuss the simplest case involving linear DNA, two digests, and no measurement error. We will refer to this problem as the double digest problem, or problem DDP. A restriction enzyme cuts a piece of DNA of length $L$ at all occurrences of a short specific pattern and the lengths of the resulting fragments are recorded. In the double digest problem we have as data the list of fragment lengths when each enzyme is used singly, say,

$$A = \{a_i: 1 \leqq i \leqq n\} \qquad \text{from the first digest}$$
$$B = \{b_i: 1 \leqq i \leqq m\} \qquad \text{from the second digest,}$$

as well as a list of double digest fragment lengths when the restriction enzymes are used in combination and the DNA cut at all occurrences specific to both patterns, say

$$C = \{c_i: 1 \leqq i \leqq n_{1,2}\};$$

only length information is retained. In general $A$, $B$, and $C$ will be multisets; that is, there may be values of fragment lengths that occur more than once. We adopt the convention that the sets $A$, $B$, and $C$ are ordered, that is, $a_i \leqq a_j$ for $i \leqq j$, and likewise for the sets $B$ and $C$. Of course

$$\sum_{1 \leqq i \leqq n} a_i = \sum_{1 \leqq i \leqq m} b_i = \sum_{1 \leqq i \leqq n_{1,2}} c_i = L,$$

since we are assuming that fragment lengths are measured in number of letters with no errors.

Given the above data the problem is to find orderings for the sets $A$ and $B$ such that the double digest implied by these orderings is, in a sense made precise below, $C$. This is a mathematical statement of the problem considered by Pearson, who solved it by exhaustive search.

We may express the double digest problem more precisely as follows. For $\sigma \in S_n$, $\mu \in S_m$ call $(\sigma, \mu)$ a configuration. By ordering A and B according to $\sigma$ and $\mu$, respectively, we obtain the set of locations of cut sites

$$S = \left\{ s\colon s = \sum_{1 \leqq j \leqq r} a_{\sigma(j)} \text{ or } s = \sum_{1 \leqq j \leqq t} b_{\mu(j)};\ 0 \leqq r \leqq n, 0 \leqq t \leqq m \right\}.$$

Since we want to record only the location of cut sites, the set $S$ is not allowed repetitions, that is, $S$ is not a multiset. Now label the elements of $S$ such that

$$S = \{ s_j\colon 0 \leqq j \leqq n_{1,2} \} \qquad \text{with } s_i \leqq s_j \text{ for } i \leqq j.$$

The double digest implied by the configuration $(\sigma, \mu)$ can now be defined by

$$C(\sigma, \mu) = \{ c_i(\sigma, \mu)\colon c_i(\sigma, \mu) = s_j - s_{j-1} \text{ for some } 1 \leqq j \leqq n_{1,2} \},$$

where we assume as usual that the set is ordered in the index $i$. The problem then is to find a configuration $(\sigma, \mu)$ such that $C = C(\sigma, \mu)$. As discussed in Section 3, this problem lies in the class of $NP$ complete problems conjectured to have no polynomial time solution.

In order to implement the simulated annealing algorithm as described above, an energy function and a neighborhood structure are required. We take as our energy function the chi-squared like criterion

$$f(\sigma, \mu) = \sum_{1 \leqq i \leqq n_{1,2}} (c_i(\sigma, \mu) - c_i)^2 / c_i;$$

note that if all measurements are error free then $f$ attains its global minimum value of zero for at least one choice $(\sigma, \mu)$.

Following Lutton and Bonomi, we define the set of neighbors of a configuration $(\sigma, \mu)$ by

$$N(\sigma, \mu) = \{(\tau, \mu)\colon \tau \in N(\sigma)\} \cup \{(\sigma, \nu)\colon \nu \in N(\mu)\},$$

where $N(\rho)$ are the neighbors used in the discussion of the travelling salesman problem above.

With these ingredients, the algorithm was tested on exact, known data from the bacteriophage lambda with restriction enzymes BamHI and EcoRI, yielding a problem size of $|A|!|B|! = 6!6! = 518{,}400$. See Daniels *et al.* [3] for the complete sequence and map information about lambda. Tempera-

ture was not lowered at the rate $c/\log(n)$ as suggested by the theorem in Geman and Geman [7], but for reasons of practicality was instead lowered exponentially. On three separate trials using various annealing schedules the solution was located after 29,702, 6895, and 3670 iterations from random initial configurations.

The algorithm was tested further on simulated data constructed by the following model. On a segment of length $n$ with sites one unit apart labeled $1, 2, \ldots, n$, assume the restriction enzyme used in the first and second single digest makes a cut at site $i$ independently with probability $p_1$, $p_2$, respectively. This model can be justified on the grounds that a segment of DNA can be approximated as a string of independent, identically distributed random variables with values in a four letter alphabet, although a first order Markov chain frequently fits real data better [18]. In addition, although in a real segment sites cut by different restriction enzymes never exactly coincide, our model allows this to occur. This feature of our model is justified by the fact that DNA segments lengths can seldom be measured precisely and that two different enzymes can cut at sites very close together. On data generated by this model the algorithm was able to locate solutions to large problems in a small number of iterations. For example, on a problem of size $(16!16!)/(2!)^7(3!)^2(4!) = 3.96 \times 10^{21}$ a solution was located in only 1635 iterations. It must be mentioned, however, that any study of the algorithms efficacy under the above probability model is confounded by the presence of multiple solutions to the exact problem in many instances. For example, a simulated problem of size 4320 was found to have 208 distinct exact solutions. This problem instance, as well as a rigorous account of the phenomenon of multiple solutions, appears in the next section. The same feature of many exact solutions must also be a property of the problem of size $3.96 \times 10^{21}$ mentioned above.

## 3. Multiplicity of Solutions in the Double Digest Problem

In many instances, the solution to the double digest problem is not unique. For example, with

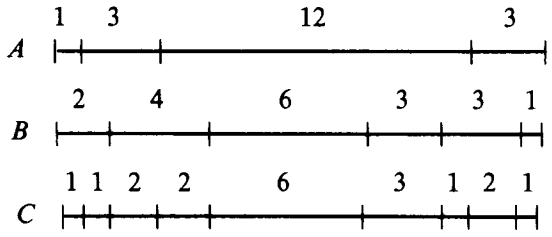
$$A = \{1, 3, 3, 12\},$$

$$B = \{1, 2, 3, 3, 4, 6\},$$

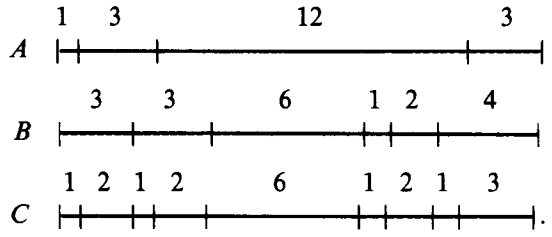
and

$$C = \{1, 1, 1, 1, 2, 2, 2, 3, 6\},$$

two distinct solutions are given by




and




In fact, the first diagram was generated as a simulated problem according to the probability model given in Section 2; the second diagram indicates the simulated annealing algorithm's output in this instance. Further investigation revealed that the problem depicted above, of size $4!6!/2!2! = 4320$, admits 208 distinct solutions. We now demonstrate that this phenomenon is far from isolated.

Below, we use the Kingman subadditive ergodic theorem to prove that the number of solutions to the double digest problem as formulated in Section 1 increases exponentially as a function of length under the probability model stated above.

For reference, we state a version of subadditive ergodic theorem here [9]. For $s, t$ nonnegative integers with $0 \leqq s \leqq t$ let $X_{s,t}$ be a collection of random variables which satisfy

(i) Whenever $s < t < u$, $X_{s,u} \leqq X_{s,t} + X_{t,u}$,

(ii) The joint distribution of $\{X_{s,t}\}$ is the same as that of $\{X_{s+1,t+1}\}$,

(iii) The expectation $g_t = E[X_{0,t}]$ exists and satisfies $g_t \geqq -Kt$ for some constant $K$ and all $t > 1$.

Then the finite $\lim_{t\to\infty} X_{0,t}/t = \lambda$ exists with probability one and in the mean.

For the probability model given above, sites labeled $1, 2, 3, \ldots$, are cut by two restriction enzymes independently with probability $p_1$, $p_2$, respectively with $p_i \in (0, 1)$.

Let a coincidence be defined to be the event that a site is cut by both restriction enzymes; such an event occurs at each site independently with probability $p_1 p_2 > 0$, and at site 0 by definition. On the sites $1, 2, 3, \ldots$, there will be an infinite number of such events. For $s, u = 0, 1, 2, \ldots$, with $0 \leqq s \leqq u$ we may consider the double digest problem for only that segment located between the $s$th and $u$th coincidence. Let $Y_{s,u}$ denote the number of solutions to the double digest problem for this segment; that is, with $A_{s,u}$, $B_{s,u}$ the sets of fragment lengths given by the first and second single digests, respectively, for only that part of the segment between the $s$th and $t$th coincidence, and $C_{s,u}$ the set of fragment lengths produced when both enzymes are used in combination for this same subsegment, $Y_{s,u}$ is the number of orderings of the sets $A_{s,u}$, $B_{s,u}$ that produce $C_{s,u}$.

It is clear that wherever $s < t < u$, given a solution for the segment between the $s$th and $t$th coincidence and a solution for the segment between the $t$th and $u$th coincidence one has a solution for the segment between the $s$th and $u$th coincidence. Hence

$$Y_{s,u} \geqq Y_{s,t} Y_{t,u}.$$

We note that the inequality may be strict as $Y_{s,u}$ counts solutions given by orderings where fragments initially between, say, the $s$th and $t$th coincidence now appear in the solution between the $t$th and $u$th coincidence. Letting

$$X_{s,t} = -\log Y_{s,t}$$

we have $s \leqq t \leqq u$ implies $X_{s,u} \leqq X_{s,t} + X_{t,u}$.

The assumption that the cuts occur independently and with equidistribution in each digest imply condition (ii) in the hypotheses of the theorem.

Last, to show condition (iii) of Kingman's theorem is satisfied, let $n_i$, $i = 1, 2, \ldots$, be the length of the segment between the $(i-1)$st and $i$th coincidence; note that $n_i$ are independent and identically distributed with $E[n_i] = 1/(p_1 p_2)$. The length of the segment from the start until the $t$th coincidence is given by $m(t) = n_1 + n_2 + \cdots + n_t$. There are $2^{(m(t)-1)}$ ways for either the first or second restriction enzyme to cut the remaining $m(t) - 1$ sites between 0 and $m(t)$, and so the total number of pairs of orderings of $A_{0,t}$, $B_{0,t}$ is bounded above by $4^{m(t)}$. Note that not all of these orderings need be solutions. Therefore

$$Y_{0,t} \leqq 4^{m(t)}$$

or

$$X_{0,t} \geqq -(\log 4) m(t)$$

so

$$E[X_{0,t}] \geqq -Kt, \qquad \text{where } K = \log(4)/p_1 p_2.$$

We may now conclude $Y_{0,t}/t \to \lambda$ with probability one. By the usual ergodic argument [9], we have $\lambda = E[\lambda]$ with probability one.

In addition, we may show that $\lambda > 0$ by the following argument. Iterating

$$Y_{s,u} \geqq Y_{s,t}Y_{t,u}$$

we obtain

$$Y_{0,t} \geqq \prod_{i=1,t} (Y_{i-1,i})$$

and so

$$E[\log(Y_{0,t})]/t \geqq E[\log(Y_{0,1})].$$

Since the example with multiple solutions depicted above has positive probability of occurring under the probability model considered,

$$P(Y_{0,1} \geqq 2) > 0.$$

This fact, together with the observation that by construction $Y_{0,1} \geqq 1$, yields $E[\log(Y_{0,1})] = \mu > 0$. Taking limits $\lambda \geqq \mu > 0$.

Letting now $Z_{m(t)}$ be the number of solutions for the segment of length $m(t)$ beginning at 0, we have by definition $Z_{m(t)} = Y_{0,t}$. Therefore

$$\lim_{t\to\infty}\log(Z_{m(t)})/m(t) = \lim_{t\to\infty}\log(Y_{0,t})/t \cdot t/m(t),$$

which, by the above and strong law of large numbers is equal to $p_1 p_2 \lambda$ with probability one. Therefore, for a segment of length $m$ we have the approximation

$$Z_m \approx \exp(\gamma m), \qquad \text{where } \gamma = p_1 p_2 \lambda;$$

that is, the number of solutions to the double digest problem increases exponentially fast as a function of the length of the segment.

## 4. Computational Complexity of the Double Digest Problem

We demonstrate below that the double digest problem is *NP* complete. It is clear that the double digest problem DDP as described above is in the class *NP*, as a nondeterministic algorithm need only guess a configuration $(\sigma, \mu)$ and check in polynomial time if $C(\sigma, \mu) = C$. The number of steps to check this is in fact linear. To show that DDP is *NP* complete we transform the partition problem to DDP.

In the partition problem, known to be *NP* complete [6], we are given a finite set $A$, say $|A| = n$, and a positive integer $s(a)$ for each $a \in A$ and

wish to determine whether there exists a subset $A' \subset A$ such that

$$\sum_{a \in A'} s(a) = \sum_{a \in A - A'} s(a).$$

If $\sum_{a \in A} s(a) = J$ is not divisible by two, there can be no such subset $A'$; else, consider as input to problem DDP the data

$$A = \{s(a_k): 1 \leqq k \leqq n\}$$
$$B = \{J/2, J/2\} \qquad \text{and set} \qquad C = A.$$

It is clear that any solution to problem DDP with this data yields a solution to the partition problem through the order of the implied digest $C$.

## 5. Circular Maps

As DNA occurs in circular as well as linear conformations, it is important to construct algorithms that can infer a map showing the location of restriction sites from sets of unordred fragment lengths when circular DNA is digested by restriction enzymes used singly and in combination. We now show how a simple modification of the algorithm described above can handle problems involving circular DNA.

We consider, as before, the double digest problem where the DNA is digested by two distinct restriction enzymes used singly and in combination; we will use the same notation as in Section 2. Here $L$ will be circumference of the DNA in base pairs. As in the linear case we are given as data three sets of unordered fragment lengths each summing to $L$, one for each time a restriction enzyme is used singly and one set of lengths when both enzymes are used together. Here, however, we are required to find circular arrangements of the single digest fragment lengths that imply the double.

As circular arrangements of fragment lengths of the two single digests may rotate relative to each other it is not enough to specify only a pair of permutations $(\sigma, \mu)$ in which the fragment lengths occur in some specified direction, counterclockwise from above, say. Distinguish then a fragment in each single digest and consider the points in the circular arrangement of each digest where these fragments are first encountered when moving counterclockwise. If $p$ measures the counterclockwise distance along the circumference of the DNA from the point in the $A$ digest to the point in the $B$ digest, then a configuration is specified by $(\sigma, \mu, p)$, where without loss of generality we may assume that $\sigma(1)$ and $\mu(1)$ correspond to the distinguished fragment in the $A$ and $B$ digest, respectively.

The implied double digest is obtained from this configuration by a reduction to the linear case as follows. We consider the digest $A$ fragments

laid out in the order dictated by $\sigma$ with the distinguished fragment as the left end. Now travel a distance $L - p$ counterclockwise from the piece that corresponds to $\mu(1)$ in the $B$ digest where we encounter the pieces in the order dictated by $\mu$. Introduce a cut at this point and using this cut as the left end of the $B$ digest align this cut with the left end of the $A$ digest pieces. The remaining fragments in the $B$ digest can now be laid out in the order dictated by $\mu$, and the double digest computed as before.

As remarked earlier, in order to implement the simulated annealing algorithm one need only specify a cost function and a neighborhood structure. We adopt the same cost function as in the linear case. Also, our notion of neighborhood here is a direct extension of that notion in the linear case; we say that two configurations are neighbors if one can be obtained from the other by reversing the order of any sequence of fragments in either digest.

As in the linear case we must consider whether any final configuration may be obtained from an initial configuration by moving through a succession of neighborhoods as those just defined. Consider first any circular arrangement of pieces of one of the single digest. Any point $e$ between two adjoining fragments may be considered momentarily as an end and the fragments may then be ordered from point $e$ in any desired permutation as described in Section 2.

To consider rotations fix point $e$ relative to the circular arrangement above. We show that the arrangement may be rotated in either direction through a distance $a_1$ along the circumference relative to the fixed point $e$. First, reverse the order of fragments between point $e$ and the fragment of length $a_1$; we include the fragment of length $a_1$ as well as the fragment with one end aligned on $e$ in the reversal so that one end of the fragment of length $a_1$ is now aligned on point $e$. Let point $f$ be the point at the end of the segment of length $a_1$ not on point $e$. Now we may let point $f$ play the role of left end and order the fragments in any desired permutation as in the linear case described in Section 2; in fact, the rotation through length $a_1$ can always be achieved in no more than four reversals. It is clear that this procedure may be used to rotate any arrangement through a distance $a_1$ in either direction.

The above process may be repeated to yield rotations through distances $k_1 a_1$ along the circumference for any integer $k_i$, and may be applied to any fragment. Rotations of size $\Sigma k_i a_i$ for any integers $k_i$ are therefore possible, and hence, rotations of any multiple of $g_1 = \gcd(a_1, a_2, \ldots, a_n)$ are achievable.

By analogous reasoning rotations of arrangements of digest $B$ through multiples of $g_2 = \gcd(b_1, b_2, \ldots, b_m)$ are attainable. Relative to each other then, the two circular arrangements may rotate through a distance $k_1 g_1 + k_2 g_2$ for any integers $k_1, k_2$, and hence, through multiples of $g = \gcd(g_1, g_2) = \gcd(a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m)$. Therefore any final con-

figuration is reachable from any initial configuration through a succession of neighborhoods when $g = 1$, and this neighborhood structure leads to an irreducible Markov chain in this case. In the (unlikely) event that $g \neq 1$ an irreducible Markov chain may be obtained by the introduction of random rotations; that is, we would say two configurations are neighbors if one can be obtained from the other by reversing the order of any sequence of fragments in one digest and then rotating either digest through any distance.

## 6. Measurement Error

Last, we consider the problem of measurement error. In real data the length of fragments are not measured precisely. In fact, measurement error is distributed (approximately) proportional to fragment length. See Schaffer [17] for discussion. Thus we are given sets of unordered fragment lengths where the sum of fragment lengths from set to set may no longer even agree.

An attempt was made to deal with this problem in the linear, double digest case again using as data the restriction map from the bacteriophage lambda with restriction enzymes BamHI and EcoRI. First, in order to simulate real measurement error as closely as possible, errors were introduced in the data by multiplying all fragment lengths by a factor of $\exp(\sigma Z)$, where $Z$ denotes a standard normal random variable independent from fragment to fragment. The fragments were then scaled so that the sum of fragment lengths in all three sets of digests were the same, just as in the case of perfect measurement. The cost function used was the same as given above. In selecting a neighbor of a given configuration first a single digest was selected, each with equal probability, and a sequence of fragments from this digest was chosen to be inverted as before. Here, however, for each fragment inverted, its length was altered randomly in the same manner that the initial measurement error was artificially introduced in the simulated problem. The algorithm would then proceed as in the case of perfect measurement as described above. This method was not successful for problems with realistic error sizes with a $\sigma$ of 0.05, say. The magnitude of success of the algorithm in the error free case, however, leads us to conjecture that the algorithm will perform well on problems with realistic error sizes if the proper notion of neighborhood were used.

## Acknowledgments

The second author is grateful to F. Blattner and D. Daniels of the University of Wisconsin Genetics Department for much valuable insight into practical and computational aspects of

restriction mapping. Both authors are grateful to Richard Arratia for assistance with the application of Kingman's ergodic theorem and to George Lueker and Roger Witney for assistance with proving DDP to be NP complete.

## REFERENCES

1. E. BONOMI, AND J-L. LUTTON, The *N*-city travelling salesman problem: Statistical mechanics and the Metropolis algorithm, *SIAM Rev.* **26** (1984) 551–568.
2. S. BREEN, M. S. WATERMAN, AND N. ZHANG, Renewal theory for several patterns, *J. Appl. Probab.* **22** (1985), 228–234.
3. D. DANIELS, J. SCHROEDER, W. SZYBALSKI, F. SANGER, A. COULSON, G. HONG, D. HILL, G. PETERSON, AND F. BLATTNER, Complete annotated lambda sequence, in "Lambda II," (R. W. Hedrix, J. W. Roberts, and F. W. Weisberg, Eds.), Cold Spring Harbor Laboratory, 1983.
4. R. DURAND AND F. BREGERERE, An efficient program to construct restriction maps from experimental data with realistic error levels, *Nucleic Acids Res.* **12**, (1985), 703–716.
5. W. M. FITCH, T. F. SMITH, AND W. W. RALPH, Mapping the order of DNA restriction fragments, *Gene* **22** (1983), 19–29.
6. M. R. GAREY AND D. S. JOHNSON, "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freeman, San Francisco, 1979.
7. S. GEMAN AND D. GEMAN, Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, (1984), 721–741.
8. B. HAJEK, Cooling schedules for optimal annealing, *Math. Oper. Res.*, in press.
9. J. F. C. KINGMAN, Subadditive ergodic theory, *Ann. Probab.* **1** (1973), 883–909.
10. S. KIRKPATRICK, C. D. GELATT, JR., M. P. VECCHI, Optimization by simulated annealing, *Science*, **220** (1983), 671–681.
11. R. LEWIN Proposal to sequence the human genome stirs debate, *Science*, **232** (1986), 1598–1600.
12. S. LIN, Computer solutions of the traveling salesman problem, *Bell System Tech. J.* **44** (1965), 2245–2269.
13. N. METROPOLIS, A. ROSENBLUTH, M. ROSENBLUTH, A. TELLER, AND E. TELLER, Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21** (1953), 1087–1092.
14. D. NATHANS AND H. O. SMITH, Restriction endonucleases in the analysis and restructuring of DNA molecules, *Ann. Rev. Biochem.* **44** (1975), 273–293.
15. C. NOLAN, G. P. MAIRNA, AND A. A. SZALAY, Plasmid mapping computer program. *Nucleic Acids Res.* **12** (1984), 717–729.
16. W. PEARSON, Automatic construction of restriction site maps, *Nucleic Acids Res.* **10** (1982), 217–227.
17. H. E. SCHAFFER, Determination of DNA fragment size from gel electrophoresis mobility, in "Statistical analysis of DNA sequence data," pp. 1–14, Dekker, New York, 1983.
18. T. F. SMITH, M. S. WATERMAN, AND J. R. SADDLER, Statistical characterization of nucleic acid sequence functional domains, *Nucleic Acids Res.* **11** (1983), 2205–2220.
19. M. STEFIK, Inferring DNA structure from segmentation data, *Artificial Intell.* **11** (1978), 85–114.
20. M. S. WATERMAN, Frequencies of restriction sites. *Nucleic Acids Res.* **11** (1983), 8951–8956.
21. M. S. WATERMAN AND J. R. GRIGGS, Interval graphs and maps of DNA, *Bull. Math. Biol.* **48**, No. 2 (1986), 189–195.
22. M. WULKAN AND T. J. LOTT, Computer aided construction of nucleic acid restriction maps using defined vectors, *Comput. Appl. Biosci.* **1** (1985), 235–239.