BIOINFORMATICS

Vol. 19 no. 11 2003, pages 1303–1310
DOI: 10.1093/bioinformatics/btg166

A comparison of physical mapping algorithms based on the maximum likelihood model

Jinling Huang[*,†] *and Suchendra M. Bhandarkar*

Department of Computer Science, University of Georgia, Athens, Georgia 30602-7404, USA

Received on October 22, 2002; revised on January 30, 2003; accepted on February 6, 2003

ABSTRACT

Motivation: Physical mapping of chromosomes using the maximum likelihood (ML) model is a problem of high computational complexity entailing both discrete optimization to recover the optimal probe order as well as continuous optimization to recover the optimal inter-probe spacings. In this paper, two versions of the genetic algorithm (GA) are proposed, one with heuristic crossover and deterministic replacement and the other with heuristic crossover and stochastic replacement, for the physical mapping problem under the maximum likelihood model. The genetic algorithms are compared with two other discrete optimization approaches, namely simulated annealing (SA) and large-step Markov chains (LSMC), in terms of solution quality and runtime efficiency.

Results: The physical mapping algorithms based on the GA, SA and LSMC have been tested using synthetic datasets and real datasets derived from cosmid libraries of the fungus *Neurospora crassa*. The GA, especially the version with heuristic crossover and stochastic replacement, is shown to consistently outperform the SA-based and LSMC-based physical mapping algorithms in terms of runtime and final solution quality. Experimental results on real datasets and simulated datasets are presented. Further improvements to the GA in the context of physical mapping under the maximum likelihood model are proposed.

Availability: The software is available upon request from the first author.

Contact: tupistra@yahoo.com

INTRODUCTION

Mapping molecular markers on chromosomes is critical for understanding the genetic structure, various functions and evolution of an organism. Chromosomal maps can be broadly categorized into genetic maps and physical maps. Genetic maps represent genetic markers in their relative order along the chromosome, where the distance between two markers is a measure of their recombination frequency and denoted by centimorgans. Although genetic maps can be used to estimate the actual physical distance between genetic markers, the result of the estimation is often not very reliable since recombination frequencies vary in different regions of a chromosome (Watson *et al.*, 1992). Physical mapping, on the other hand, determines the actual physical locations of molecular markers on a chromosome. Physical maps are often of much higher resolution and the distance between two markers in a physical map is measured by the number of intervening nucleotide base pairs (Brody *et al.*, 1991). The highest resolution map for a chromosome is the complete nucleotide sequence of that chromosome. Consequently, the physical map is a powerful tool to isolate and manipulate genes and to study the organization and evolution of genome (Prade *et al.*, 1997).

Physical mapping can be accomplished by ordering distinguishable DNA fragments (clones or contigs) derived from a library of cloned DNA fragments according to their positions in the genome. This can be done using a variety of techniques that are specific to an experimental protocol and type of data collected such as nonunique probes mapping (Alizadeh *et al.*, 1995), unique probes mapping (Alizadeh *et al.*, 1994; Jain and Myers, 1997), unique endprobes mapping (Christof *et al.*, 1997), restriction fragments mapping (Fasulo *et al.*, 1997; Jiang and Karp, 1998), radiation-hybrid mapping (Ben-Dor and Chor, 1997; Slonim *et al.*, 1997) and optical mapping (Muthukrishnan and Parida, 1997; Lee *et al.*, 1998). The cloned fragments, once mapped, can be further cut by restriction enzymes into smaller DNA fragments which are sequenced and assembled to yield the complete sequence of the genome (Kececioglu and Myers, 1995).

Our method of reconstructing the physical ordering of probes/clones along the chromosome is based on a maximum likelihood (ML) model proposed by Shete (1998) and Kececioglu *et al.* (2000) and also described in Bhandarkar *et al.* (2001). The ML model derives a likelihood function for the parameters underlying the

*To whom correspondence should be addressed.
† Present address: Center for Tropical and Emerging Global Diseases, 623 Biological Science Building, University of Georgia, Athens, GA 30602, USA.

problem to be solved and seeks to obtain the parameter values that maximize the statistical likelihood of resulting in the observed data. The ML model is not only intuitive, but also well grounded in statistical theory (Hogg and Craig, 1995). The ML model has been successfully used in the chromosome mapping of *Aspergillus nidulans* as a part of the Fungal Genome Initiative at the University of Georgia and consistently provided better results than other existing procedures (Shete, 1998).

SYSTEMS AND METHODS

The physical mapping technique adopted in our project is based on the *sampling without replacement* protocol (Fu *et al.*, 1992). This protocol has been used successfully in the physical mapping of several fungal organisms, including *Aspergillus nidulans* (Prade *et al.*, 1997), *Schizosaccharomyces pombe* (Mizukami *et al.*, 1993) and *Pneumocystis carinii* (Arnold and Cushion, 1997). Under this protocol, a genomic DNA library L consisting of many overlapping clones of equal length and covering the entire chromosome is first created. The probe set P and clone set C are then chosen iteratively. During the ith iteration, a new clone is selected at random from the library L and designated as the ith probe P_i. This newly selected probe P_i is hybridized against all the remaining clones in L. A clone is removed from the library L if it has a positive reaction with P_i. The clone/probe hybridization reactions are recorded in a binary hybridization matrix H. If the jth clone has a positive reaction with the ith probe, then $H_{ji} = 1$; else $H_{ji} = 0$. The above steps are repeated until the library L becomes empty.

It can be shown that the resulting probe set P is essentially a maximal set of non-overlapping clones spanning the length of the chromosome (Kececioglu *et al.*, 2000). If the probes in the probe set P can be ordered with respect to their positions along the chromosome, then by examining the binary hybridization signatures of the clones in the probe/clone hybridization matrix, the clones linking successive probes in the ordering can also be inferred. The ordered probe set P and the linking clones can be further used to reconstruct a minimal set of overlapping probes and clones (i.e. the minimum tiling) that spans the length of the chromosome. The minimum tiling in conjunction with the sequencing of each individual clone/probe followed by a sequence assembly procedure can then be used to reconstruct the DNA sequence of the entire chromosome (Kececioglu and Myers, 1995). In practice, the reconstruction procedure is rarely error free. The most common form of error arises from a false hybridization signature. This occurs when a clone and a probe have a false positive hybridization reaction and H_{ij} is encoded as 1 when, in fact, it should be 0, or conversely, when a clone and a probe have a false negative hybridization reaction and H_{ij} is encoded as 0 when it should be 1 (Bhandarkar *et al.*, 2001).

The ML estimation procedure determines the optimal ordering of probes Π in the probe set P and the optimal inter-probe spacings Y under a probabilistic model of hybridization errors due to false positives and false negatives. Once the optimal probe ordering is determined, the ordering of clones can be obtained by examining the probe/clone hybridization matrix H.

The negative log-likelihood (NLL) function derived from the ML model can be expressed as

$$f(\Pi, Y) = C - \sum_{i=1}^{k} \ln\Bigg\{ R_i - \sum_{j=1}^{n+1} (a_{i,\pi_j} - 1) \times (a_{i,\pi_{j-1}} - 1) \min(Y_j, \mathrm{M}) \Bigg\} \quad (1)$$

where C is a constant given by

$$C = k \ln(N - M) - P \ln \frac{\rho}{(1-\rho)} - nk \ln(1-\rho), \quad (2)$$

$$R_i = N - nM + M \sum_{j=1}^{n-1} (a_{i,\pi_j} a_{i,\pi_j+1}) \quad (3)$$

$$a_{i,j} = \begin{cases} \frac{\eta}{(1-\rho)} & \text{if} \quad H_{i,j} = 0 \quad \text{and} \quad j = 1, \ldots, n \\ \frac{(1-\eta)}{\rho} & \text{if} \quad H_{i,j} = 1 \quad \text{and} \quad j = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

and where

N is the length of the chromosome;

M is the length of a clone/probe;

n is the number of probes selected from the library;

k is the number of clones in the library;

ρ is the probability of a false positive hybridization;

η is the probability of a false negative hybridization;

$H = ((H_{i,j}))_{1 \leqslant i \leqslant k, 1 \leqslant j \leqslant n}$ is the clone/probe hybridization matrix;

H_{ij} is the hybridization result of the ith clone with the jth probe;

$\Pi = (\pi_1, \ldots, \pi_n)$ is a permutation of $\{1, 2, \ldots, n\}$;

$Y = (Y_1, Y_2, \ldots Y_n)$ is the inter-probe spacing vector where Y_i is the spacing between P_{π_i} and $P_{\pi_{i-1}}$, Y_1 is

the spacing between the start of the chromosome and the beginning of the first probe P_{π_1}, and Y_{n+1} is the spacing between the end of the last probe P_{π_n} and the end of the chromosome.

Optimization of the ML function entails the minimization of the NLL function $f(\Pi, Y)$, which leads to the probe ordering and inter-probe spacings that maximize the probability of occurrence of the observed hybridization matrix H. Minimization of the NLL function entails a two-tier approach. At the higher tier, optimization methods such as simulated annealing (SA), large-step Markov chains (LSMC), or the genetic algorithm (GA) can be used to optimize the NLL function with respect to the discrete parameter Π (i.e. the probe ordering). At the lower tier, a gradient descent search procedure can be used to optimize the NLL function with respect to the continuous parameter Y (i.e., the inter-probe spacings) for a given probe ordering Π (Bhandarkar *et al.*, 2001). The conjugate gradient descent search procedure used in our project is a refinement of the straightforward steepest descent search procedure. In the conjugate gradient descent search procedure, the gradient vector is projected on a set of mutually orthonormal (i.e. conjugate) direction vectors. The minimization along each conjugate direction can proceed independently of the others. The conjugate gradient descent search procedure is known to have a convergence rate that is faster than that of the straightforward steepest descent search procedure (Kincaid and Cheney, 1991).

SA and LSMC algorithms: SA is an iterative optimization approach analogous to the process of gradual cooling of a physical system. A typical SA algorithm starts with a given temperature and an initial solution (or state). The temperature is then gradually decreased according to an annealing function. Each temperature value corresponds to an annealing step that consists of several iterations where each iteration comprises of three phases as described below:

(a) *Perturb*: A perturbation is generated by reversing the probe order within a randomly chosen block of probes in the current ordering Π_i to yield a new candidate probe ordering Π_j (Bhandarkar *et al.*, 2001). This perturbation is referred to as a *2-opt* perturbation in the context of the traveling salesman problem (TSP).

(b) *Evaluate*: The NLL function value $f(\Pi, Y)$ for the new probe ordering Π_j is computed. This is achieved by searching for the optimal inter-probe spacing Y_j for Π_j using the conjugate gradient descent search procedure and then computing the NLL function value under Π_j and Y_j.

(c) *Decide*: If $f(\Pi_j, Y_j) < f(\Pi_i, Y_i)$, then Π_j is accepted as the new candidate probe ordering; otherwise Π_j is accepted as the new candidate probe ordering with a probability p computed using the Metropolis function

$$p = \exp\left(-\frac{f(\Pi_j, Y_j) - f(\Pi_i, Y_i)}{T_i}\right)$$

at temperature T_i whereas Π_i is retained with probability $(1-p)$. In our implementation, a random number is generated using a pseudorandom number generator with a uniform distribution in the range $[0, 1]$. If this random number happens to be less than p, then Π_j is accepted as the current probe ordering; otherwise Π_i is retained.

For a given temperature value T_i (or annealing step), a sufficient number of *Perturb–Evaluate–Decide* cycles leads to an equilibrium resulting in a stationary Boltzmann distribution of solution states (Geman and Geman, 1984). Therefore, a sufficient number of iterations should be run to approach the equilibrium. At higher temperatures, since almost any change to the current candidate solution state can be potentially accepted, uphill moves that result in an increase in the NLL function value are highly probable and SA resembles a completely random search. This allows large-scale exploration of the search space while avoiding being trapped in a local optimum. At lower temperatures, SA resembles a deterministic local search since uphill moves are less probable.

The LSMC algorithm (Fig. 1) combines the Metropolis decision function with an exhaustive local search using the *2-opt* perturbation. The current solution at every stage is guaranteed to be locally optimal under the *2-opt* perturbation. In the perturb phase, the current solution (which is locally optimal) is subject to a non-local perturbation termed as a *double-bridge kick* (Martin *et al.*, 1991) which results in a transition to a non-local point in the search space. Using the *2-opt* perturbation, an exhaustive local search is performed starting from this new point resulting in a new local optimum. The choice between the new local optimum and the current solution is then made using the Metropolis decision function as in the case of SA.

The exhaustive local search using the *2-opt* perturbation would make LSMC computationally extremely intensive since the NLL function value would need to be evaluated at each step. As an effective compromise, the exhaustive local search is performed using a modified objective function. The modified objective function $f_D(\Pi, Y) = f_D(\Pi)$ computes the sum of the Hamming distances between the binary hybridization signatures of successive probes in a given probe ordering Π. The column of the hybridization matrix H corresponding to a probe is deemed to be the binary hybridization signature of that probe. The exhaustive local search seeks the local

Choose a random order of probes Π and a sufficiently high value of the temperature T.
Perform an exhaustive local search using the *2-opt* perturbation and the Hamming distance objective function f_D on Π to yield a new probe order Π'. Compute $f(\Pi', \hat{Y}_{\Pi'})$ using the conjugate gradient descent search algorithm.

While (not converged)
{
  for ($i = 1$; $i <$ Max_iterations; $i = i + 1$)
  {
    (a) Perform a non-local *double-bridge* perturbation on Π' to yield a new probe order Π''.
    (b) Perform an exhaustive local search using the *2-opt* perturbation and the Hamming distance objective function f_D on Π'' to yield a new probe order Π'''. Compute $f(\Pi''', \hat{Y}_{\Pi'''})$ using the conjugate gradient descent search algorithm.
    (c) If $f(\Pi''', \hat{Y}_{\Pi'''}) < f(\Pi', \hat{Y}_{\Pi'})$
        Replace the existing solution $(\Pi', \hat{Y}_{\Pi'})$ by the new solution $(\Pi''', \hat{Y}_{\Pi'''})$.
      Else
        Generate a random number x that is uniformly distributed in the interval [0, 1].
        If $x < \exp(-(f(\Pi''', \hat{Y}_{\Pi'''}) - f(\Pi', \hat{Y}_{\Pi'}))/T)$
          Replace the existing solution $(\Pi', \hat{Y}_{\Pi'})$ by the new solution $(\Pi''', \hat{Y}_{\Pi'''})$.
        Else
          Retain the existing solution $(\Pi', \hat{Y}_{\Pi'})$.
  }
  Reduce the temperature T using the annealing schedule $T = A(T)$.
  Check for convergence.
}

Fig. 1. LSMC algorithm for computing the optimal probe ordering and optimal inter-probe spacings in the context of ML model-based physical mapping.

minimum of $f_D(\Pi)$. Since the modified objective function $f_D(\Pi)$ is much easier to compute than the original NLL function $f(\Pi, Y)$, the exhaustive local search procedure is very fast. Because LSMC samples only the space of locally optimal solutions, it is often computationally more efficient than SA, which attempts to sample the entire search space (Bhandarkar *et al.*, 2002).

Genetic algorithm: The GA is an adaptive optimization approach modeled on the process of natural selection underlying biological evolution. Unlike SA and LSMC that start with a single candidate solution, the GA pursues an optimal solution by starting from an initial ensemble or population of candidate solutions, and iterating through generations of candidate solutions for a globally optimal solution. The final solution is chosen from an ensemble of locally optimal solutions.

We have designed two versions of the GA in the context of the ML model-based physical mapping problem. The first version (Fig. 2) combines the stochastic decision function used in SA and LSMC, whereas the second version is based strictly on deterministic search for new candidate solutions. In both versions, the population is initialized by a series of *double-bridge* perturbations followed by exhaustive local searches using the *2-opt* perturbation.

A heuristic crossover operator, originally proposed by Jog *et al.* (1989) in the context of the TSP, was used in our GA. We model the probe order as a tour where the probes represent the nodes (i.e. cities) and the distance between probes is the Hamming distance between their corresponding binary hybridization signatures. The heuristic crossover operator is described as follows:

(a) Select two parental chromosomes.

(b) Choose a start node from one of the chromosomes selected for crossover.

(c) Compare the two edges emanating from the start node in the two parental chromosomes and choose the shorter edge; if the shorter edge leads to an illegal tour, choose the other edge; if both edges

Initialize population P of chromosomes using *double-bridge* perturbations followed by exhaustive local search using the *2-opt* perturbation.

$T = T\_\max$.
while (not converged)
{
  for ($i = 1$; $i <$ Population_size; $i = i + 1$)
  {
    (a) Select two parents using the roulette wheel selection procedure.
    (b) Apply heuristic crossover operator to the selected parents to create an offspring S.
    (c) Perform an exhaustive local search using the *2-opt* perturbation starting from S and identify the locally optimal solution S^*.
    (d) Perform conjugate gradient descent search and compute the NLL function value at S^*.
    (e) Perform a mutation using a *double-bridge* perturbation with probability *prob* on S^* followed by an exhaustive local search using the *2-opt* perturbation.
    (f) Compute *f_delta*, the change in the NLL value between the less fit parent and S^*.
    (g) Retain the less fit parent with the probability p computed using the Boltzmann function.
  }
  Update *prob*.
  Update the temperature $T = A(T)$.
  Check for convergence.
}
Select the best individual from the population as the globally optimal solution.

Fig. 2. GA with stochastic replacement for ML model-based physical mapping.

introduce illegal tours, choose an edge from the remaining nodes that represents the shortest distance from the start node.

(d) Choose the new node as the start node and repeat step (c) until a complete tour is generated.

Parental chromosomes are selected using the roulette wheel selection procedure (Goldberg, 1989) followed by application of the heuristic crossover with a slight modification. Note that if the selected start node is close to the end of the probe order, recombination will not be effective in most cases where the two parents are similar. For example, if the start node is selected at a point after which the probe orderings associated with both parental chromosomes are the same, no actual exchange of parental chromosomal segments will occur in spite of the application of the heuristic crossover. To avoid this scenario, we start from the beginning of the probe order whenever a node close to the end of the probe order is selected as the start node. Mutation, in our algorithms, is implemented using the non-local *double-bridge* perturbation followed by an exhaustive local search using the *2-opt* perturbation. The mutation rate is set dynamically based on the genetic variation retained in the population. In the case of the GA that uses stochastic replacement, the less fit parent is retained with the probability p computed using the Boltzmann decision function $p = \frac{1}{1+e^{(E_i - E_j)/T}}$ where E_i and E_j are the NLL function values associated with the parent chromosome and child chromosome, respectively, and T is the temperature parameter which is updated using an annealing function as in SA and LSMC. For the GA that uses deterministic replacement, the better offspring always replaces the less fit parent.

EXPERIMENTAL RESULTS

The algorithms were implemented on a shared-memory symmetric Multiprocessor (SMP) consisting of four 700 MHz Intel Xeon processors with 1 MB cache per processor and 1 GB of shared memory and running the Solaris-x86 operating system. However, since the code is serial, only a single processor in the SMP was used. The number of chromosomes in the population for both GA versions was set to 10. For the GA based on stochastic replacement (Fig. 2) as well as the SA and LSMC algorithms, the initial value of temperature T was chosen to be 1. The temperature was systematically reduced using a geometric

annealing schedule of the form $T_{\text{next}} = \alpha.T_{\text{prev}}$, with the annealing factor $\alpha = 0.9$. Both false positive and negative rates were assumed to be 2%, which is consistent with the physical mapping experiments conducted in the Department of Genetics at the University of Georgia. In the case of the GA, the annealing process was terminated when no improvement was detected for all the individuals in the population in two successive generations. In the case of the SA and LSMC algorithms, the annealing process was terminated when the same solution was returned by two successive annealing steps.

The ML model-based physical mapping algorithms based on the GA, SA and LSMC were tested using artificially created datasets with $(n, k) = (50, 300)$, (100, 650) and (200, 1300) (Shete, 1998), and real data sets derived from cosmid libraries *cosmid2* ($n = 109$, $k = 2046$) and *cosmid3* ($n = 111$, $k = 1937$) of the fungus *Neurospora crassa*, which were made available to us by Dr. Jonathan Arnold, Department of Genetics, University of Georgia. The artificial datasets were generated by first generating an artificial chromosome of length N which was represented by the interval $[0, N]$ on the real axis. A set of clones, each of length M, that are uniformly distributed along the length of the chromosome was generated via uniform sampling of the interval $[0, N - M]$ for the left end of the clone. The number of clones was chosen to ensure at least a 5-fold coverage of the chromosome. The probe set and the hybridization matrix were generated by simulating the *sampling without replacement* protocol on the aforementioned clone set. Finally, false positive and false negative hybridization errors were introduced into the hybridization matrix with predetermined error rates (Shete, 1998).

Comparison of execution time and final values of the NLL function between the two GA versions shows similar results on the synthetic dataset with number of probes $n = 50$. The GA that used deterministic replacement yielded a NLL function value of 1548.2667 in 3 334 seconds and the GA that used stochastic replacement yielded a value of 1548.3959 in 3 525 seconds. For the larger synthetic dataset with number of probes $n = 100$ and real datasets *cosmid2* and *cosmid3*, the GA that used stochastic replacement yielded lower NLL function values for all the datasets, suggesting thereby that the algorithm based on deterministic replacement is more likely to be trapped in a local optimum.

Table 1 compares the execution time and NLL function values resulting from SA, LSMC and the GA with stochastic replacement. For the same dataset, the GA almost always had a solution with lower NLL value than LSMC and SA. The only exception is in the case of the real dataset *cosmid3*, where the GA yielded a slightly higher NLL function value (less than 1% difference) than LSMC but with a much shorter execution time (less than half).

Table 2 compares the number of probe suborderings (i.e. contigs) recovered by SA, LSMC, and the GA with stochastic replacement on the synthetic datasets $(n, k) = (50, 300)$, (100, 650) and (200, 1300) where the true probe order is a single contig $1, 2, 3, \ldots, n$. In an ideal case, the optimization algorithm should able to recover the true probe order as a single contig, but this is very unlikely in practice due to the presence of false hybridization signatures and artifacts. In a more realistic scenario the physical mapping algorithm would be expected to recover probe suborderings that could then be manually manipulated (via translation and probe order reversal) to yield the final probe order. The fewer and longer these probe suborderings or contigs, the less intensive the subsequent manual editing in order to recover the desired probe ordering. For our test datasets, the GA consistently yielded fewer and longer probe suborderings (contigs) than SA and LSMC, suggesting a better solution quality in the case of the GA with stochastic replacement.

DISCUSSION

Tests using both synthetic and real datasets have shown that the GA with heuristic crossover and with either deterministic or stochastic replacement is superior to SA and LSMC in terms of quality of the solution to the ML model-based physical mapping problem. In almost all cases, the GA results in solutions with lower NLL function values as well as fewer and longer contigs within a reasonable time frame. Therefore the GA with heuristic crossover represents a major improvement over existing optimization techniques such as SA and LSMC in terms of solution quality in the context of the ML model-based physical mapping problem.

The GA adopted in this study uniquely combines the strengths of heuristic crossover that yields better solutions (i.e. probe suborderings) from parental chromosomes, non-local perturbation that results in distinct solution states, and exhaustive local search that ensures an ensemble of locally optimal solutions in each generation. A key criterion underlying the choice of the crossover operator is that a better solution to the ML model-based physical mapping problem should, in most cases, have a low Hamming distance-based objective function value as well. This assumption was first tested in the exhaustive local search procedure in the LSMC-based physical mapping algorithm and yielded some very promising results compared to its SA-based counterpart (Bhandarkar *et al.*, 2002). Most available crossover operators are random assortments of the parental chromosomes (Goldberg, 1989; Michalewicz, 1994). These crossover operators, in spite of their ability to bring more genetic variation into the population, are not particularly useful since the resulting offspring are often less fit than their parents. As a

Table 1. Comparisons between the GA with stochastic replacement, SA and LSMC in terms of run time and solution quality

Data	GA time (sec)	GA value	LSMC time (sec)	LSMC value	SA time (sec)	SA value
$n = 50$	3 525	1 548.39	10 746	1 624.09	15 076	1 665.37
$n = 100$	10 919	4 262.80	7 459	4 297.50	6 265	4 288.64
$n = 200$	181 311	11 159.48	105 893	11 515.13	31 013	11 574.75
cosmid2	27 962	12 731.37	34 704	12 757.55	108 499	12 949.99
cosmid3	14 922	12 584.62	30 138	12 501.88	45 533	13 212.85

$n = 50$ ($k = 300$), $n = 100$ ($k = 650$), and $n = 200$ ($k = 1300$) are artificial data sets.
cosmid2 ($n = 109$, $k = 2046$) and *cosmid3* ($n = 111$, $k = 1937$) are derived from the cosmid libraries of the fungus *Neurospora crassa*.

Table 2. Number of contigs recovered by the GA with stochastic replacement, LSMC and SA

Number of probes	GA	LSMC	SA
50	4	12	12
100	9	14	13
200	9	24	27

result, most of these offspring are immediately eliminated from the population even though a significant amount of time has been spent creating and evaluating them. This scenario is particularly detrimental for problems of high computational complexity such as the ML model-based physical mapping since the computational cost associated with the evaluation of each new probe ordering is significant. The heuristic crossover operator adopted in our GA extends the probe ordering of the offspring by intentionally selecting better probe suborderings from two parental chromosomes, which also are locally optimal solutions. In most cases, and especially during the early stages of the search process where the genetic variation in the population is greater, this approach will produce offspring with lower Hamming distance-based objective function values than those of their parents.

The initial population in GA is created by a series of *double-bridge* perturbations followed by exhaustive local searches using the *2-opt* perturbation. Since the *double-bridge* is a non-local perturbation technique, its application will yield a set of distinct solution states that are not reachable from each other via a series of local perturbations. The local exhaustive search, on the other hand, ensures that each chromosome in the initial population represents a locally optimal solution to the problem under the *2-opt* perturbation. In each generation, an exhaustive local search is also applied to the new offspring to ensure that all individuals in the population represent locally optimal solutions to the problem at a certain stage in the search process. This strategy is useful since locally optimal solutions under the *2-opt* perturbation provide better starting points from which to further search for a global optimum than randomly created offspring. Although the computational cost associated with the local exhaustive search using the *2-opt* perturbation is substantial for some large datasets, it is still insignificant compared to the cost associated with the conjugate gradient descent search procedure and evaluation of the NLL function value. Maintaining a population of locally optimal solutions over successive generations essentially avoids the significant computational cost incurred by selecting random, less fit individuals for reproduction. In the absence of the exhaustive local search procedure, the convergence of the GA is often about three times slower and leads to solutions that are significantly inferior (Huang and Bhandarkar, unpublished data).

The convergence time of our implementation of the GA compares well with that of SA and LSMC, but is a candidate for further improvement. In our implementation of the GA, searching for a globally optimal solution terminates when no improvement has been detected for all individuals in the population in two successive generations. Due to the nature of the interaction among individuals in the population via heuristic crossover, this termination condition is rather stringent and will most likely exhaust the evolutionary potential of the population and lead to the homogenization and convergence of the entire population to a global optimum. For smaller populations, since convergence to a globally optimal solution is a relatively rapid process, this condition is easily satisfied. When a larger population size is used, this termination condition can lead to an extremely long process of homogenization and convergence even though a globally optimal solution may have already found its way into the current population. Alternatively, it would be more proper to terminate the search when the best individual in the population remains unchanged for a specified number of successive iterations.

ACKNOWLEDGEMENTS

The research was supported in part by a grant from the US Department of Agriculture (USDA) under the NRI Competitive Grants Program.

REFERENCES

Alizadeh,F., Karp,R.M., Weisser,D.K. and Zweig,G. (1994) Physical mapping of chromosomes using unique probes. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM Press, New York, NY, pp. 489–500.

Alizadeh,F., Karp,R.M., Newberg,L.A. and Weisser,D.K. (1995) Physical mapping of chromosomes: a combinatorial problem in molecular biology. *Algorithmica*, **13**, 52–76.

Arnold,J. and Cushion,M.T. (1997) Constructing a physical map of the *Pneumocystis* genome. *J. Eukaryot. Microbiol.*, **44**, 8S.

Ben-Dor,A. and Chor,B. (1997) On constructing radiation hybrid maps. In *Proceedings of the First ACM Conference on Computational Molecular Biology*. ACM Press, New York, NY, pp. 17–26.

Bhandarkar,S.M., Machaka,S.A., Shete,S.S. and Kota,R.N. (2001) Parallel computation of a maximum-likelihood estimator of a physical map. *Genetics*, **157**, 1021–1043.

Bhandarkar,S.M., Huang,J. and Arnold,J. (2002) Parallel Monte Carlo methods for physical mapping of chromosomes. In *Proceedings of IEEE Computer Society Bioinformatics Conference*. IEEE Press, Los Alamitos, CA, pp. 64–75.

Brody,H., Griffith,J., Cuticchia,A.J., Arnold,J. and Timberlake,W.E. (1991) Chromosome-specific recombinant DNA libraries from the fungus *Aspergillus nidulans*. *Nucleic Acids Res.*, **19**, 3105–3109.

Christof,T., Junger,M., Kececioglu,J., Mutzel,P. and Reinelt,G. (1997) A branch-and-cut approach to physical mapping of chromosomes by unique end-probes. *J. Comput. Biol.*, **4**, 433–447.

Fasulo,D.P., Jiang,T., Karp,R.M., Settergren,R. and Thayer,E.C. (1997) An algorithmic approach to multiple complete digest mapping. In *Proceedings of the First ACM Conference on Computational Molecular Biology*. ACM Press, New York, NY, pp. 118–127.

Fu,Y.X., Timberlake,W.E. and Arnold,J. (1992) On the design of genome mapping experiments using short synthetic oligonucleotides. *Biometrics*, **48**, 337–359.

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.

Hogg,R.V. and Craig,A.T. (1995) *Introduction to Mathematical Statistics*, 5th edn, Prentice-Hall, Englewood Cliffs, NJ.

Jain,M. and Myers,E.W. (1997) Algorithms for computing and integrating physical maps using unique probes. *J. Comput. Biol.*, **4**, 449–466.

Jiang,T. and Karp,R.M. (1998) Mapping clones with a given ordering or interleaving. *Algorithmica*, **21**, 262–284.

Jog,P., Suh,J.Y. and Gucht,D.V. (1989) The effects of population size, heuristic crossover and local improvement on a genetic algorithm for the traveling salesman problem. In *Proceedings of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers, Los Altos, CA, pp. 110–115.

Kececioglu,J.D. and Myers,E.W. (1995) Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, **13**, 7–51.

Kececioglu,J.D., Shete,S.S. and Arnold,J. (2000) Reconstructing distances in physical maps of chromosomes with nonoverlapping probes. In *Proceedings of the Fourth ACM Conference on Computational Molecular Biology*. Tokyo, Japan, pp. 183–192.

Kincaid,D. and Cheney,W. (1991) *Numerical Analysis*. Brooks/Cole, Pacific Grove, CA.

Lee,J.K., Dancik,V. and Waterman,M.S. (1998) Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. In *Proceedings of the Second ACM Conference on Computational Molecular Biology*. ACM Press, New York, NY, pp. 147–152.

Martin,O., Otto,S.W. and Felten,E.W. (1991) Large-step Markov chains for the traveling salesman problem. *Complex Systems*, **5**, 299–326.

Michalewicz,Z. (1994) *Genetic Algorithms + Data Structures = Evolution Programs*, 2nd edn, Springer, Berlin, Germany.

Mizukami,T., Chang,W.I., Garkavtsev,I., Kaplan,N., Lombardi,D. *et al.* (1993) A 13 Kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell*, **73**, 121–132.

Muthukrishnan,S. and Parida,L. (1997) Towards constructing physical maps by optical mapping: an effective, simple, combinatorial approach. In *Proceedings of the First ACM Conference on Computational Molecular Biology*. ACM Press, New York, NY, pp. 209–219.

Prade,R.A., Griffith,J., Kochut,K., Arnold,J. and Timberlake,W.E. (1997) In vitro reconstruction of the *Aspergillus nidulans* genome. *Proc. Natl Acad. Sci. USA*, **94**, 14564–14569.

Shete,S.S. (1998) *Estimation problems in physical mapping of a chromosome and in a branching process with immigration*, Ph.D. dissertation, Department of Statistics, University of Georgia, Athens, GA.

Slonim,D., Kruglyak,L., Stein,L. and Lander,E. (1997) Building human genome maps with radiation hybrids. *J. Comput. Biol.*, **4**, 487–504.

Watson,J.D., Gilman,M., Witkowski,J. and Zoller,M. (1992) *Recombination DNA*, 2nd edn, Scientific American Books, New York, NY.