BIOINFORMATICS

Vol. 17 no. 3 2001
Pages 205–213

Physical mapping with automatic capture of hybridization data

David Hall[1, 2], Suchendra M. Bhandarkar[1,*], Jonathan Arnold[2] and Tongzhang Jiang[1]

[1]Department of Computer Science and [2]Department of Genetics, The University of Georgia, Athens, GA 30602, USA

Received on June 23, 2000; revised on October 25, 2000; accepted on November 20, 2000

ABSTRACT

Motivation: Contig maps are a type of physical map that show the native order of a set of overlapping genomic clones. Overlaps between clones can be detected by finding common sequences using a number of experimental protocols including hybridization of probes. All current mapping algorithms of which we are aware require that hybridizations be scored using a fixed number of discrete values (typically 0/1 or high/medium/low). When hybridization data is captured automatically using digital equipment, this provides the opportunity for hybridization intensities to be used in map construction. More fine-grained distinctions in the levels of hybridization may be exploited by algorithms to generate more accurate physical maps.

Results: We describe an approach to creating contig maps that uses measured hybridization intensities instead of data scored with a fixed number of discrete values. We describe and compare four algorithms for creating physical maps with hybridization intensities. Simulations using measured intensities sampled from actual data on *Aspergillus nidulans* indicate that using hybridization intensities rather than data that is automatically scored with respect to threshold values may yield more accurate physical maps.

Availability: All software programs described in this paper may be obtained by contacting the authors.

Contact: suchi@cs.uga.edu

INTRODUCTION

A contig map is a type of physical map that gives a partial ordering of a set of overlapping genomic clones. Contig maps aid in the positional cloning of genes, serve as a framework for genome sequencing (Chumakov *et al.*, 1995; McPherson, 1997), and can be used to study the large-scale organization of genomes (Prade *et al.*, 1997). Several types of assays are used to detect clone overlaps. One that has been widely used is based on fingerprinting clones by treating them with restriction enzymes and then determining the size of the resulting fragments (Coulson *et al.*, 1986; Olson *et al.*, 1986). An advantage with this protocol is that chimeric clones can often be identified (McPherson, 1997). Gel electrophoresis is commonly used to measure the size of restriction fragments. A shortcoming with electrophoresis is its inability to resolve fragments that are nearly the same size. Optical mapping technology has provided a solution to this problem (Cai *et al.*, 1995). However, the equipment costs are considerable. Algorithms for constructing maps using restriction digestion fingerprinting data are described in Coulson *et al.* (1986), Carrano *et al.* (1989), Stallings *et al.* (1990), and Gillett *et al.* (1995).

Another fingerprinting method is based on hybridization of clones to oligonucleotide probes (Lehrach, 1990). Algorithms for recovering the ordering of clones with such data are described in Cuticchia *et al.* (1992), Fu *et al.* (1992), and Mayraz and Shamir (1999). However, this protocol has seen only limited use in actual mapping projects (Mayraz and Shamir, 1999). Two mapping protocols that have been more widely used are based on the detection of unique sequences within clones (Hudson *et al.*, 1995). These are mapping by hybridization to unique probes (Torney, 1990) and Sequence-Tagged Site (STS) content mapping (Green and Olson, 1990). Algorithms for map construction with these data types usually focus on ordering the probes or STSs. Once they are ordered, a contig map can be created by overlaying clones on the ordered markers using the hybridization or STS content data (Mott *et al.*, 1993). The problem of ordering probes or STSs is difficult because of errors in detecting the sequences as well as the possible existence of chimeric clones (Greenberg and Istrail, 1995).

The same algorithmic approaches can be used to order unique probes or STSs. For the purpose of brevity we will use the term *marker* to denote either of these. Most algorithms use a binary system for encoding exper-

*To whom correspondence should be addressed.

imental results. Marker content data are encoded in a matrix A, where a_{ij} is 1 if the ith clone is believed to contain the jth marker, or 0 otherwise. One class of algorithms for ordering markers using such data is based on solving the Traveling Salesman Problem (TSP). A quasi-distance metric between markers, such as Hamming distance (Cuticchia *et al.*, 1992), is defined, and an ordering of markers is sought that minimizes the sum of distances between adjacent markers. Typically a local search algorithm, such as simulated annealing, is used to minimize the distance-based objective function. Traveling salesman algorithms for ordering markers are described in Cuticchia *et al.* (1992), Mott *et al.* (1993), Wang *et al.* (1994), and Bhandarkar and Machaka (1997). Another class of algorithms for binary data is based on the consecutive 1's property. If there is no experimental error and the columns of A are permuted so they correspond to the true marker order, then in each row of the matrix, all 1's will occur consecutively (Greenberg and Istrail, 1995). Algorithms based on the consecutive 1's property are described in Greenberg and Istrail (1995), Alizadeh *et al.* (1995), Jain and Myers (1997), Christoff *et al.* (1997), and Christoff and Kececioglu (1999). Approaches for binary data based on minimal spanning trees (Mott *et al.*, 1993; Nadkarni *et al.*, 1996) and genetic algorithms (Tsai and Kao, 2000) have also been described. Kececioglu *et al.* (2000) describe a maximum likelihood model for ordering markers and estimating the distance between them using binary data.

A method for simultaneously building a contig map and sequencing the clones has been described by Venter *et al.* (1996). This method relies on sequence alignment and has several advantages over the methods described above. Clones that minimally overlap can be identified, so a minimum number of clones covering the chromosome can be selected. There is no need to create a sequence-ready map prior to sequencing, as sequencing and mapping are carried out simultaneously. Improved sequence assembly algorithms developed at Celera Genomics (http://www.celera.com) purportedly eliminate the need to even create a sequence-ready contig map. Entire chromosomes can be shotgun sequenced and then assembled. Such algorithms at this time are however proprietary.

Despite the development of these new methods for generating contig maps and the entire chromosome sequence, the older mapping methods based on restriction digestion, hybridization to probes, and STS content provide relatively inexpensive approaches to rapidly create physical maps when sequence information is not needed. A hybridization-based mapping protocol is advantageous in that a high degree of parallelism can be achieved. DNA from thousands of clones can be fixed to a solid surface (e.g. a nylon filter). The clones can all be probed in parallel with the same radioactively or fluourescently labeled probe. If the clones have been partitioned into chromosome-specific libraries, then each filter can be probed with multiple probes from different chromosomes simultaneously to achieve a speedup proportional to the number of chromosomes.

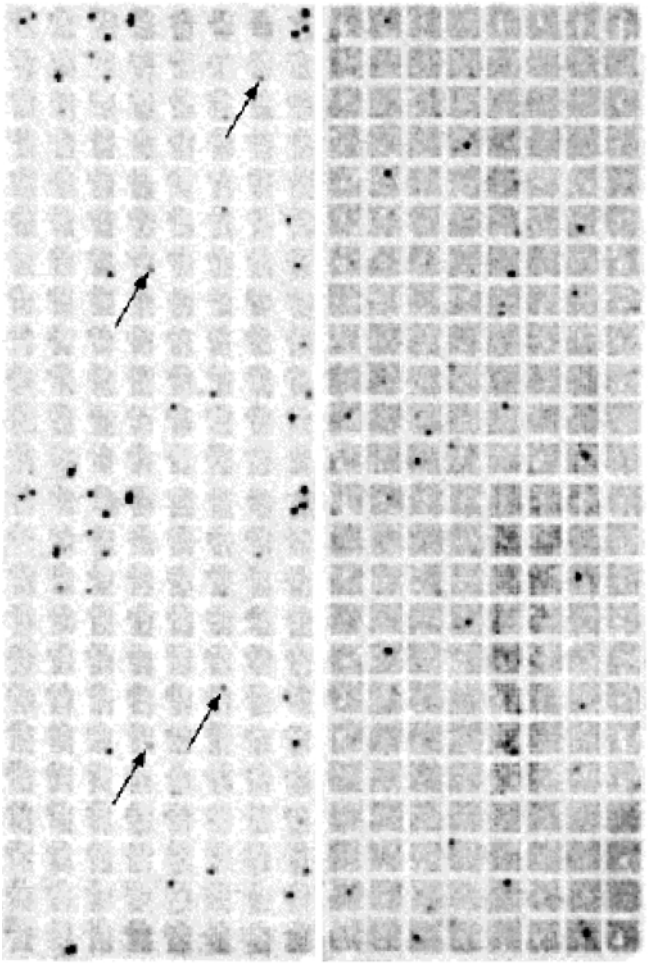

Fig. 1. An autoradiograph of two radioactively probed filters each containing DNA from 3072 clones. Examples of hybridizations of intermediate intensity are indicated by arrows.

Hybridization data from a filter can be automatically scored by software by applying a threshold to raw hybridization intensities measured from digitally captured images of the filter. Intensities greater than or equal to the threshold are scored as positive hybridizations, and intensities below the threshold are scored as negative. Figure 1 shows an autoradiograph of two radioactively probed filters. It can be seen that there are hybridizations that clearly should be scored as positive and hybridizations that clearly should be scored as negative. However, there are also hybridizations of intermediate value (examples are indicated by arrows in Figure 1). When scoring by thresholding, these intermediate intensities would be converted

to 1 or 0. This may be a source of experimental error.

It has been recognized that allowing hybridization experiments to be scored with more than two values (e.g. high, medium, low, non-hybridizing) may be advantageous (Soderlund and Dunham, 1995; Sasinowska and Sasinowski, 1999). Physical mapping algorithms based on data that can have more than two discrete values have been described in two papers that we are aware of. The program SAM (Soderlund and Dunham, 1995) accepts as input, hybridization data that are scored as high, medium, or low. The program computes a weight between each pair of markers and uses a stochastic optimization algorithm to place pairs of markers with large weights in adjacent positions. The weight between two markers is incremented by a value for each clone that hybridizes to both markers. This value is a function of the hybridization intensity.

Sasinowska and Sasinowski (1999) describe a clone ordering algorithm that uses hybridization data that can be scored with an arbitrary number of discrete values. (In their paper they used four values, 0, 3, 6, and 9.) The weight between two clones is equal to the scalar product of the vectors encoding the hybridization data for the clones, i.e. the weight between clones c_i and c_j is equal to the scalar product of a_i and a_j, the hybridization signatures of the clones. Contigs are built using a greedy algorithm and a novel statistical measurement of robustness for data consisting of multiple discrete values.

This paper explores an alternative to scoring with a system of discrete values. We describe two new algorithms for creating contig maps where the data may take any value over a range (e.g. [0, 1]). One is a TSP-based algorithm for ordering clones or probes. The other is a matrix-based algorithm for ordering unique probes when clones and probes are of the same size. The latter case arises in the *sampling without replacement* mapping protocol used in the mapping of *Schizosaccharomyces pombe* (Mizukami *et al.*, 1993) and *Aspergillus nidulans* (Prade *et al.*, 1997) and is currently being used to map the genome of *Pneumocystis carinii* (Arnold and Cushion, 1997), and *Nectria haematococca* (Enkerli *et al.*, 2000). We also describe a modification of the algorithm of Sasinowska and Sasinowski (1999) for ordering probes using hybridization intensities. This algorithm uses an objective function based on the scalar product of vectors and a greedy search method. We further modify this algorithm by substituting a stochastic search algorithm in place of the greedy algorithm. The performance of these algorithms is evaluated against other algorithms for scored data. Simulations with hybridization intensities samples from actual mapping data from *A.nidulans* (Prade *et al.*, 1997) suggest that using hybridization intensities rather than automatically scored data may yield more accurate physical maps.

SYSTEM AND METHODS

Hybridization experiments were carried out as follows. DNA from genomic libraries of Lambda-based cosmid clones were stamped onto 5 × 3.5 inch nylon membranes containing DNA from 3072 clones. Hybridization was carried out using ^{32}P-labeled probes. Probes were derived from the genomic library itself. Hybridization was detected by autoradiography.

Autoradiographs were digitized by a scanner. The resulting TIFF format files were converted into an 8-bit grayscale raster format using the UNIX graphics utility *xv*. Hybridization intensities for each clone in the image were computed by a program called AUTOREAD, that was developed internally. AUTOREAD provides the same basic functionality as commercial array reading applications, such as ArrayVision (http://imagingresearch.com). The program was written in C++ and compiled with the GNU *g++* compiler on a Sun Enterprise 250 computer running Solaris 7. The program generates a text file containing a matrix of intensities, each of which corresponds to a clone in the image. Another C++ program then processes these files for input into the mapping programs as follows. The negatives of the values in the text file are computed by subtracting each value from 255 (the largest number that can be represented with 8 bits). Negatives are made so that strong hybridization intensities correspond to large values. All values from the filter are then normalized to the range [0, 1] as follows. The value *background* is computed by finding the average value of all pixels in the image. The value $max\_intensity$ is computed by finding the clone with greatest average grayscale value. The normalized hybridization intensity *norm* for a clone is computed by the following formula,

$$\text{norm} = \max\left\{0, \frac{\text{value} - \text{background}}{\text{max\_intensity} - \text{background}}\right\}$$

where *value* denotes the raw hybridization intensity for the clone. The output of this program is a text file containing a matrix of normalized values.

All programs written to test the algorithms in this paper were written in C++ and compiled and executed on a Sun Enterprise 250 computer running Solaris 7.

ALGORITHMS

We first describe a TSP objective function that can be used in the ordering of either clones or unique probes. Without loss of generality we will describe the objective function in terms of ordering probes. Let A be a matrix where a_{ki} gives the hybridization intensity measured between clone c_k and probe p_i. We first define a quasi-distance metric between two probes, p_i and p_j. (Note: this metric can also

be used between clones.)

$$D(p_i, p_j) = \sum_k |a_{ki} - a_{kj}|.$$

Informally, the distance between probes is equal to the sum of the magnitude of the differences in hybridization intensities between the probes. If the data are binary, then this metric is the Hamming distance. A TSP-based objective function is now presented. Let $P^\pi = \langle p_1^\pi, p_2^\pi, \ldots, p_n^\pi \rangle$ denote a probe permutation.

$$\text{objective}(P^\pi) = \sum_{i=1}^{n-1} D(p_i^\pi, p_{i+1}^\pi). \quad (1)$$

By minimizing this objective, a permutation of probes is found such that the sum of distances defined by the above metric between all pairs of adjacent probes is a minimum.

A matrix-based objective function for ordering unique probes is now described for the case where the probes and clones are of the same size. If this holds, and the probes are non-overlapping, then each clone should hybridize to at most two probes. If the data are noise-free (i.e. only the large hybridization intensities correspond to clones and probes that truly overlap), and if the columns of A are permuted so that they correspond to the correct ordering of probes, then in each row there should be at most two large values in adjacent positions. Let A^π denote the matrix A where the columns are permuted to correspond to the probe permutation P^π. We now define the second objective function as:

$$\text{objective}(P^\pi) = \sum_i \max_j \{a_{ij}^\pi + a_{i(j+1)}^\pi\}. \quad (2)$$

By maximizing this objective function we seek to place large values in the rows of A in adjacent positions. If the data are binary, a consecutive 1's matrix would maximize this objective function. For both functions (1) and (2) we used the microcannonical annealing algorithm (Creutz, 1983; Bhandarkar and Machaka, 1997) to search for a minimum or maximum value, respectively. This algorithm, with the parameters used in this study, is shown in Figure 2.

A modified version of the Sasinowska and Sasinowski (1999) algorithm that uses hybridization intensities is described. The Sasinowska and Sasinowski algorithm begins by selecting an initial 'comparison' clone. Contigs are built iteratively by selecting the unmapped clone with greatest weight relative to the comparison clone. This new clone then becomes the comparison clone for the next iteration. The weight between clones c_i and c_j is the scalar product of the vectors representing the two clones (i.e. rows a_i and a_j of A). The original algorithm also uses a statistical measurement of robustness based on multiple discrete hybridization values. If fingerprinting data are available, these are utilized by the algorithm as well. Our modification consisted of dropping the measurement of robustness and not accepting restriction fingerprinting data. We will refer to this algorithm as SP-GREEDY (scalar product with a greedy search). We also evaluated an algorithm based on the scalar product but which uses microcannonical annealing (Creutz, 1983) rather than a greedy search. We will refer to this as SP-MCA. Table 1 summarizes the features and differences of all algorithms evaluated in this study.

```
1. for i ← 1 to N
2.         for j ← 1 to N
3.                 E[i][j] ← E_max
4. V ← initial value of objective function
5. while not finished begin
6.         for i ← 1 to MAXCOUNT begin
7.                 Randomly choose p and q such that 1 ≤ p < q ≤ N
8.                 Reverse block of probes between and including p and q
9.                 ΔV ← change in objective function
10.                Accept permutation if one of the following is true
11.                        (i) ΔV < 0
12.                        (ii) 0 ≤ ΔV ≤ E[p][q]
13.                if permutation is accepted begin
14.                        E[p][q] = E[q][p] = E[p][q] - ΔV
15.                        V ← ΔV
16.                endif else
17.                        Restore previous probe order
18.        endfor
19.        for i ← 1 to N
20.                for j ← 1 to N
21.                        E[i][j] = E[i][j] × factor
22.        if V is unchanged for K iterations halt
23. endwhile
```

Fig. 2. The microcannonical annealing algorithm used in this study. N is the number of probes. Parameters that were used are as follows: $E_{\max} = 0.5$, factor $= 0.5$, $K = 3$, MAXCOUNT $= 100 \times N$. Note that as written this algorithm will find the permutation that minimizes the objective function. To find the permutation maximizing the objective function, the direction of the relational operators in lines 11 and 12 should be reversed. (e.g. $\leqslant$ becomes $\geqslant$).

IMPLEMENTATION

We evaluated algorithms (1), (2), SP-GREEDY, and SP-MCA on simulated clones that use hybridization intensities sampled from actual data. Data were generated by sampling hybridization intensities from the *Aspergillus*

Table 1. Summary of features of all the algorithms evaluated in this study. The abbreviation MCA denotes the microcannonical annealing search algorithm

Algorithm	Objective function	Search method
Algorithm (1)	Equation (1)	MCA
Algorithm (2)	Equation (2)	MCA
SP-GREEDY	Sum of scalar products	Greedy
SP-MCA	Sum of scalar products	MCA
HD-MCA	Sum of hamming distances	MCA
ML-MCA	Maximum posterior probability	MCA

Table 2. Percentage of correct probe adjacencies recovered by algorithms (1), (2), SP-GREEDY, and SP-MCA on simulated genomes with hybridization intensities sampled from data from *A.nidulans* (Prade *et al.*, 1997). Mean values for 100 data sets are reported

Algorithm	Probe adjacencies recovered	
	Coverage = 5 (%)	Coverage = 15 (%)
(1)	76	99
(2)	80	99
SP-GREEDY	66	97
SP-MCA	75	99

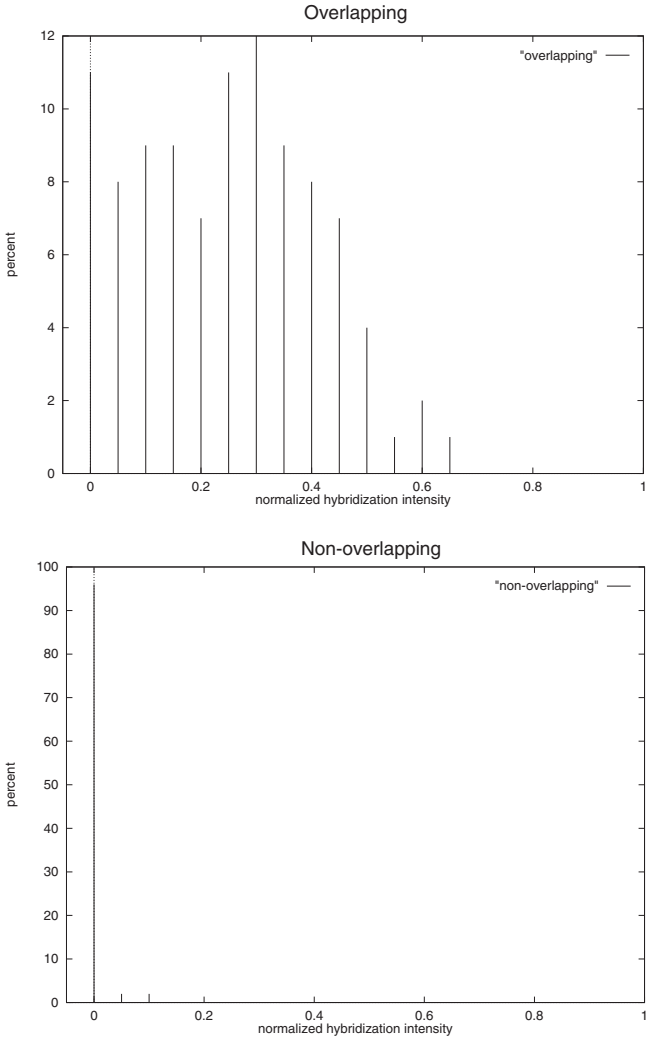

Fig. 3. Distribution of normalized hybridization intensities for clone and probe pairs that overlap, and for pairs that do not overlap. The cardinality of these sets was 250 and 2500, respectively. Data were sampled from *A.nidulans* physical mapping data (Prade *et al.*, 1997). Values in each of these sets were rounded to the next smallest 0.05 for plotting.

mapping project (Prade *et al.*, 1997). The source of these intensities consisted of autoradiographs of radioactively probed filters. The method used to capture intensities from the autoradiographs is discussed in the section System and methods. All intensities are in the range $[0, 1]$. Hybridization intensities were sampled from clone–probe pairs known to overlap or not to overlap. Let O denote the set of intensities from overlapping clones and probes and NO denote the set from non-overlapping pairs. The cardinality of these sets was 250 and 2500, respectively. The distribution of values for these two sets is shown in Figure 3. Artificial data were generated to resemble data from the mapping of a small chromosome (e.g. from a typical fungus) via the sampling without replacement protocol (Prade *et al.*, 1997). First, clones and probes were simulated *in silico*. Clones of 40 kb in size were generated by a Poisson process to produce either a coverage of 5 or 15, for a 2 Mb chromosome. A maximal set of non-overlapping clones was chosen from among all the clones and designated as the set of probes. For each clone–probe pair a hybridization intensity was generated. For simulated clone–probe pairs that overlapped, this was done by selecting a value at random from O with replacement. For non-overlapping simulated clone–probe pairs the value was selected randomly from NO with replacement.

One hundred independent data sets were generated in this manner. The quality of the resulting probe orderings is reported in terms of the percentage of correct probe adjacencies recovered by the algorithms. Results are shown in Table 2. It can be seen that at a coverage of 15, there is little difference in the performance of these algorithms. Nearly 100% of probe adjacencies are recovered on average for each algorithm. However, at a coverage of 5 there are clear differences. Algorithm (2) performed the best (80% of probe adjacencies recovered on average). Algorithms (1) and SP-MCA performed similarly (about 75% of probe adjacencies recovered on average). The SP-GREEDY algorithm performed the worst (66% of probe adjacencies recovered on average).

The best and worst performing algorithms above were compared to two algorithms that use binary scored data, the ODS algorithm of Cuticchia *et al.* (1992) and the maximum likelihood algorithm of Alizadeh *et al.* (1995). The ODS algorithm is based on Hamming distance and simulated annealing. The maximum likelihood algorithm of Alizadeh *et al.* uses simulated annealing and a weighted sum of error objective function based on the consecutive 1's property. For each of these algorithms, microcannonical annealing (Creutz, 1983) was used in place of simulated annealing as it was shown to find as good a solution as simulated annealing an order of magnitude faster (Bhandarkar and Machaka, 1997). We denote these algorithms as HD-MCA and ML-MCA, respectively. Data were scored by software for each algorithm. The same data sets that were used in Table 2 were also used here. Hybridization intensities greater than or equal to a threshold value were scored as 1, and intensities below the threshold were scored as 0. Several different thresholds were evaluated. The results from these simulations are shown in Figure 4. The results from the best and worse performing algorithm from Table 2 are also plotted for comparison. It can be seen that at a coverage of 5 the choice of a threshold value is very important. At a threshold value of 0.15, HD-MCA and ML-MCA both recovered close to 75% of correct probe adjacencies on average. At a threshold of 0.05 and 0.25, these algorithms recovered around 60% of correct probe adjacencies on average. Neither recovered as many adjacencies on average at the best threshold value tested than algorithm (2), which again recovered 80% of adjacencies. Algorithms HD-MCA and ML-MCA performed better than SP-GREEDY at a threshold of 0.15. At a coverage of 15, HD-MCA and ML-MCA perform about as well as algorithm (2) and SP-GREEDY at threshold values 0.05, 0.15, and 0.25.

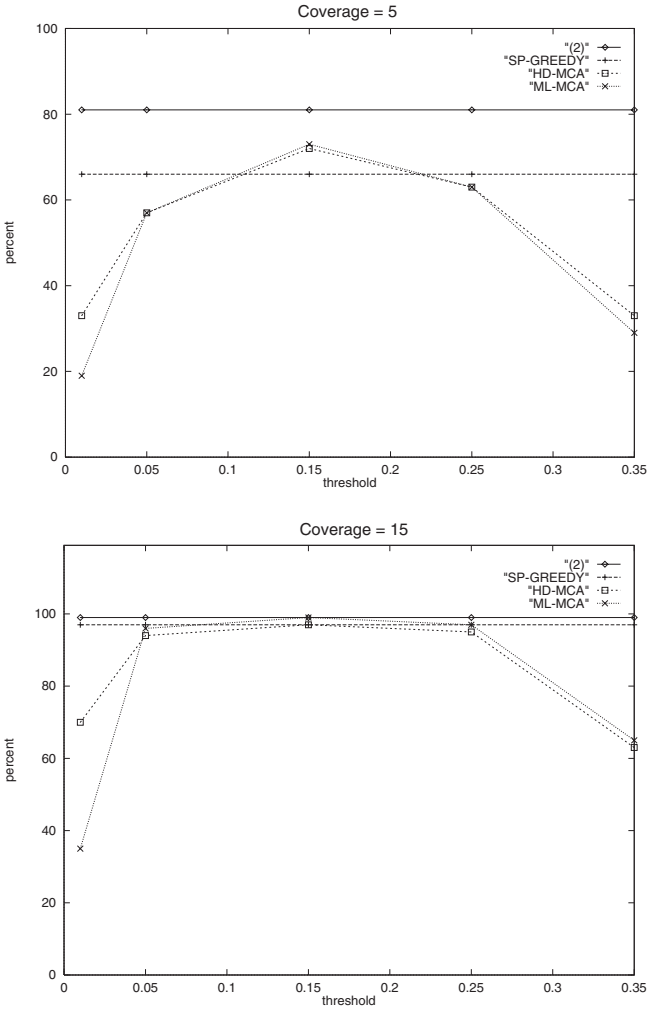

Fig. 4. Percentage of correct probe adjacencies recovered by algorithms (2), SP-GREEDY, HD-MCA, and ML-MCA on simulated genomes with hybridization intensities sampled from data from *A.nidulans* (Prade *et al.*, 1997). For algorithms HD-MCA, and ML-MCA hybridization intensities were converted to binary numbers by applying a threshold. Each point represents the mean value of 100 independent data sets.

The Sasinowska and Sasinowski (1999) algorithm accepts data scored with multiple discrete values. This algorithm first builds contigs and then orders probes based on the clone ordering. All of the other algorithms examined in this study order probes first. Recall that the SP-GREEDY algorithm is similar to the Sasinowska and Sasinowski algorithm in that the objective function is based on the scalar product of vectors and that probes are ordered greedily. In evaluating the authors' program, we found that the software predicts clone overlaps very accurately. However, the SP-GREEDY algorithm better predicted the order of probes (data not shown). In order to fairly compare the authors' general approach with the other algorithms in this report, we evaluated the SP-GREEDY algorithm on data scored with multiple discrete values, instead of the authors' original program. In order to ascertain if using hybridization intensities instead of data scored with multiple discrete values has an impact on map quality, the data sets used in Table 2 were scored by software to generate data with 4 possible values, 0, 3, 6, and 9. Two thresholds were used. Intensities greater than 0, but less than the lower threshold were scored as 3. Intensities greater than or equal to the lower threshold but less than the upper threshold were scored as 6. Intensities greater than or equal to the upper threshold were scored as 9. Intensities of 0 were scored as 0. Algorithms SP-GREEDY and SP-MCA were run on these data. These sets of points are labeled DISCRETE-SP-GREEDY and DISCRETE-SP-MCA in Figure 5, respectively. Three different ratios of lower threshold to upper threshold were evaluated, 0.25, 0.5, and 0.75 (data not shown). A lower

threshold that was half of the upper threshold gave the best results. These results are presented in Figure 5. Results from these algorithms on hybridization intensities which are reported in Table 2 are also plotted for comparison. It can be seen in this figure that for both algorithms at a coverage of 5, a greater percentage of correct probe adjacencies are recovered when hybridization intensities are used rather than discrete data. At a coverage of 15 the difference is less, as is the case in Figure 4. As in Table 2, SP-MCA performs better than SP-GREEDY.

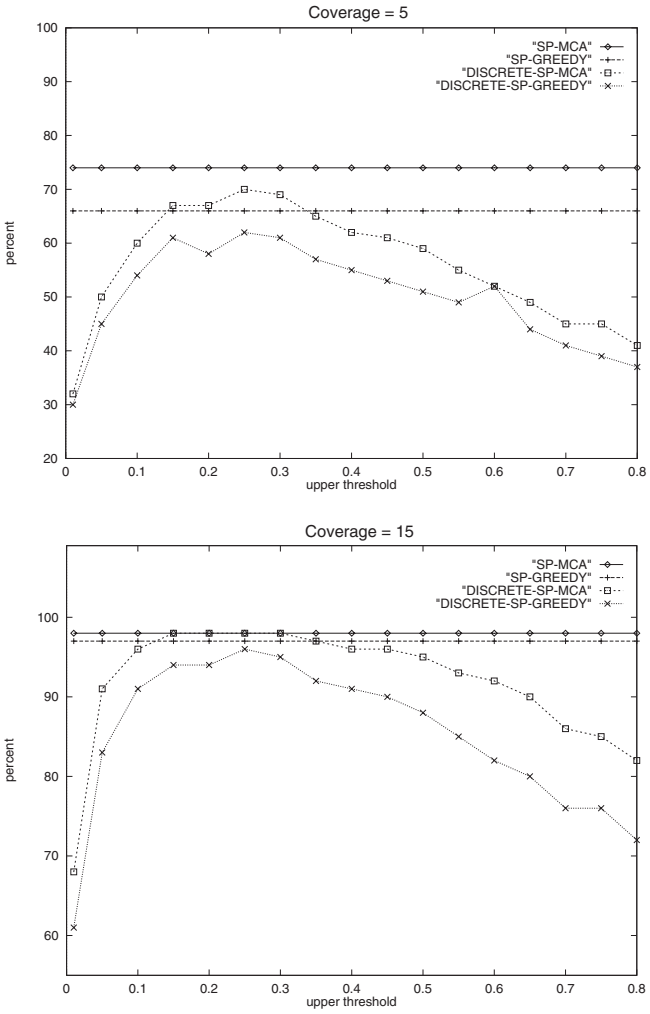

Fig. 5. Percentage of correct probe adjacencies recovered by algorithms SP-GREEDY and SP-MCA on simulated genomes with hybridization intensities sampled from data from *A.nidulans* (Prade *et al.*, 1997). The data points labeled SP-GREEDY and SP-MCA result from ordering probes using hybridization intensities. The data points labeled DISCRETE-SP-GREEDY and DISCRETE-SP-MCA are from results using data that has been scored with four possible values produced by applying two threshold values. The lower threshold was half the value of the upper threshold. Each points represents the mean of 100 independent data sets.

DISCUSSION

There are likely a number of factors that can attenuate the measured intensity of clone–probe hybridization, including variations in the amount of DNA on the filter from different clones, and intrinsic features of certain sequences. Traditionally, hybridization data has been recorded using a binary system. With binary scored data, errors in scoring (i.e. false positive and false negative errors) contribute to the difficulty of constructing physical maps. Some algorithms (Alizadeh *et al.*, 1995; Jain and Myers, 1997; Christoff *et al.*, 1997; Christoff and Kececioglu, 1999; Kececioglu *et al.*, 2000) explicitly model the problem of false positive and false negative errors. Allowing hybridization data to be recorded with more than two values enables hybridization values of intermediate intensity to be more 'safely' used in map construction. For instance, assume that a particular hybridization of intermediate intensity that would be scored as 1 under a binary system takes some value between 0 and 1 instead. Furthermore, assume that this hybridization places the clone in particular map location. If one or more strong hybridizations are observed that place the clone in another map position, then it can be reasonably assumed that the later is the correct placement. Likewise, in the absence of strong hybridization of a clone to any probe, intermediate hybridization intensities may indicate where to place the clone in the map. Under binary scoring such potentially useful information on hybridization intensity is lost.

Scoring hybridizations as low, medium, and high (Soderlund and Dunham, 1995; Sasinowska and Sasinowski, 1999) is a very reasonable approach for capturing information about hybridization intensity when data scoring is done manually. In a high-throughput environment it is desirable to automate time-consuming and labor-intensive tasks such as capturing data. Automation also makes data capture less subjective and more consistent. Data can be based on a measurable quality, such as the intensity of labeling at locations on an autoradiograph. Thus, differences in hybridization intensity can be more finely distinguished than when data are recorded using a system of multiple discrete values.

We describe in this paper an approach to physical mapping by hybridization that does not require the scoring of data using discrete values. Simulations carried out in this study suggest this may be a better approach in the context of automatic data capture than approaches based on scoring of data. We described two new algorithms for ordering probes using hybridization intensities. We also described modifications to an algorithm described in Sasinowska and Sasinowski (1999) that allow hybridization intensities to be used. Simulations suggest that algorithm (2)

performs the best out of these (Table 2). On these data, when the stochastic search algorithm microcannonical annealing (Creutz, 1983) was used with a scalar product-based objective function instead of a greedy search algorithm, a greater percentage of probe adjacencies was recovered (Table 2, SP-GREEDY versus SP-MCA). In all simulations, the greatest number of adjacencies were recovered when hybridization intensities were used instead of scored data. The same algorithms (SP-GREEDY and SP-MCA) even recovered more probe adjacencies when the data consisted of hybridization intensities, rather than discrete values (Figure 5).

In Figures 4 and 5 the range of good threshold values increases from a coverage of 5 to a coverage of 15. At the best threshold value the algorithms that use discrete values do relatively well as compared to the algorithms that use hybridization intensities. The best threshold value for any data set would certainly depend on the distribution of intensities, and may not be easy to determine in practice, especially at low coverage. This consideration along with the results from the simulations in this study suggest that creating physical maps using hybridization intensities rather than scored data may be a better approach in the context of automatic data capture.

ACKNOWLEDGEMENTS

This research was funded in part by an NRICGP grant from the US Department of Agriculture and in part by Microbial Genetics Grant MCB-9630910 from NSF.

REFERENCES

Alizadeh,F., Karp,R.M., Weisser,D.K. and Zweig,G. (1995) Physical mapping of chromosomes using unique probes. *J. Comp. Biol.*, **2**, 159–184.

Arnold,J. and Cushion,M.T. (1997) Constructing a physical map of the *Pneumocystis* genome. *J. Euk. Microbiol.*, **6**, 8S.

Bhandarkar,S.M. and Machaka,S. (1997) Chromosome reconstruction from physical maps using a cluster of workstations. *J. Supercomput.*, **11**, 61–86.

Cai,W., Hiroyuzi,A., Stanton,V.P., Housman,D.E., Wang,Y. and Schwartz,D.C. (1995) Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl Acad. Sci. USA*, **92**, 5164–5168.

Carrano,A.V., Lamerdin,J., Ashworth,L.K., Watkins,B., Bascomb,E., Slezak,T., Raff,M., De Jong,P.J., Keith,D., McBride,L., Meister,S. and Kronick,M. (1989) A high-resolution, fluorescence-based, semiautomated method for DNA fingerprinting. *Genomics*, **4**, 129–136.

Christoff,T. and Kececioglu,J. (1999) Computing physical maps of chromosomes with nonoverlapping probes by branch and cut. *Proceedings of the 3rd ACM Conference on Computational Molecular Biology,* pp. 115–123.

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

Chumakov,I.M., Rigault,P., Le Gall,I., Bellanne-Chantelot,C., Billault,A., Guillou,S., Soularue,P., Guasconi,G., Poullier,E., Gros,I., Belova,M., Sambucy,J.-L., Susini,L., Gervy,P., Gilber,F., Beaufils,S., Bui,H., Massert,C., De Tand,M.-F., Dukasz,F., Lecoulant,S., Ougen,P., Perrot,V., Saumier,M., Soravito,C., Bahouayila,R., Cohen-Akenine,A., Barrilot,E., Bertrand,S., Codani,J.-J., Caterina,D., Georges,I., Lacroix,B., Lucotte,G., Sahbatou,M., Schmit,C., Sangourard,M., Tubacher,E., Dib,C., Faure,S., Fizames,C., Gyapay,G., Millasseau,P., Nguyen,S., Muselet,D., Vignal,A., Morissette,J., Menninger,J., Lieman,J., Desai,T., Banks,A., Bray-Ward,P., Ward,D., Hudson,T., Gerety,S., Foote,S., Stein,L., Page,D.C., Lander,E.S., Weissenbach,J., Le Paslier,D. and Cohen,D. (1995) A YAC contig map of the human genome. *Nature*, **377** (suppl.), 175–297.

Coulson,A., Sulston,J., Brenner,S. and Karn,J. (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **83**, 7821–7825.

Creutz,M. (1983) Microcannonical Monte Carlo simulation. *Phys. Rev. Lett.*, **50**, 1411–1414.

Cuticchia,A.J., Arnold,J. and Timberlake,E. (1992) The use of simulated annealing in chromosome reconstruction experiments based on binary scoring. *Genetics*, **132**, 591–601.

Enkerli,J., Reed,H., Briley,A., Bhatt,G. and Covert,S.F. (2000) Physical map of a conditionally dispensable chromosome in *Nectria haematococca* MP VI and location of chromosome breakpoints. *Genetics*, in press.

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

Gillett,W., Daues,J., Hanks,L. and Capra,R. (1995) Fragment collapsing and splitting while assembling high-resolution restriction maps. *J. Comp. Biol.*, **2**, 185–205.

Green,E.D. and Olson,M.V. (1990) Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc. Natl Acad. Sci. USA*, **87**, 1213–1217.

Greenberg,D.S. and Istrail,S. (1995) Physical mapping by STS hybridization: algorithmic strategies and the challenge of software evaluation. *J. Comp. Biol.*, **2**, 219–273.

Hudson,T.J., Stein,L.D., Gerety,S.S., Ma,J., Castle,A.B., Silva,J., Slonim,D.K., Baptista,R., Kruglyak,L., Xu,S.-H., Hu,X., Colbert,A.M.E., Rosenberg,C., Reeve-Daly,M.P., Rozen,S., Hui,L., Wu,X., Vestergaard,C., Wilson,K.M., Bae,J.S., Maitra,S., Ganiatsas,S., Evans,C.A., DeAngelis,M.M., Ingalls,K.A., Nahf,R.W., Horton,L.T., Anderson,M.O., Collymore,A.J., Ye,W., Kouyoumijian,V., Zemsteva,I.S., Tam,J., Devine,R., Courtney,D.F., Renaud,M.T., Nguyen,H., O'Connor,T.J., Fizames,C., Faure,S., Gyapay,G., Dib,C., Morissette,J., Orlin,J.B., Birren,B.W., Goodman,N., Weissenbach,J., Hawkins,T.L., Foote,S., Page,D.C. and Lander,E.S. (1995) An STS-based map of the human genome. *Science*, **270**, 1945–1954.

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

Kececioglu,J., Shete,S. and Arnold,J. (2000) Reconstructing order and distance in physical maps using nonoverlapping probes. *Proceedings of the 4th ACM Conference on Computational Molecular Biology (RECOMB 2000),* pp. 81–89.

Lehrach,H. (1990) *Genetic and Physical Mapping*. Davies,K.E. and Tilghman,S.M. (eds), CSH Press, Plainview, NY.

Mayraz,G. and Shamir,S. (1999) Construction of physical maps from oligonucleotide fingerprints data. Computing physical maps of chromosomes with nonoverlapping probes by branch and cut. In *Proceedings of the 3rd ACM Conference on Computational Molecular Biology,* pp. 268–277.

McPherson,J.D. (1997) Sequence ready—or not? *Genome Res.*, **7**, 1111–1113.

Mizukami,T., Chang,W.I., Garkatseve,I., Kaplan,N., Lombardi,D., Matsumoto,T., Niwa,O., Kounosu,A., Yanagida,M., Marr,T.G. and Beach,D. (1993) A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell*, **73**, 121–132.

Mott,R., Grigoriev,A., Maier,E., Hoheisel,J. and Lehrach,H. (1993) Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucleic Acids Res.*, **21**, 1965–1974.

Nadkarni,P.M., Banks,A., Montgomery,K., LeBlanc-Stracewski,J., Miller,P. and Krauter,K. (1996) *Genomics*, **31**, 301–310.

Olson,M.V., Dutchik,J.E., Graham,M.Y., Brodeur,G.M., Helms,C., Frank,M., MacCollin,M., Scheinman,R. and Frank,T. (1986) Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA*, **83**, 7826–7830.

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**, 14,564–14,569.

Sasinowska,H. and Sasinowski,M. (1999) An algorithm for the assembly of robust physical maps based on a combination of multi-level hybridization data and fingerprinting data. *Comput. Chem.*, **23**, 251–262.

Soderlund,C.A. and Dunham,I. (1995) SAM: a system of iteratively building marker maps. *Comput. Appl. Biosci.*, **6**, 645–655.

Stallings,R.L., Torney,D.C., Hildebrand,C.E., Longmire,J.L., Deaven,L.L., Jett,J.H., Dogett,N.A. and Moyzis,R.K. (1990) Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl Acad. Sci. USA*, **87**, 6218–6222.

Torney,D.C. (1990) Mapping using unique sequences. *J. Mol. Biol.*, **217**, 259–264.

Tsai,H. and Kao,C. (2000) Using genetic algorithms to construct physical maps of chromosomes with unique probes. In Miyano,S., Shamir,R. and Takagi,T. (eds), *Currents in Computational Molecular Biology.* Universal Academy Press, Tokyo, Japan, pp. 167–168.

Venter,J.C., Smith,H.O. and Hood,L. (1996) A new strategy for genome sequencing. *Nature*, **381**, 364–366.

Wang,Y., Prade,R.A., Griffith,J., Timberlake,W.E. and Arnold,J. (1994) A fast random cost algorithm for physical mapping. *Proc. Natl Acad. Sci. USA*, **91**, 11,094–11,098.