Pp. 395–408

Contig Selection in Physical Mapping

STEFFEN HEBER,[1,2] JENS STOYE,[1] MARCUS FROHME,[2] JÖRG HOHEISEL,[2]
and MARTIN VINGRON[1]

ABSTRACT

In physical mapping, one orders a set of genetic landmarks or a library of cloned fragments of DNA according to their position in the genome. Our approach to physical mapping divides the problem into smaller and easier subproblems by partitioning the probe set into independent parts (probe contigs). For this purpose we introduce a new distance function between probes, the *averaged rank distance* (ARD) derived from bootstrap resampling of the raw data. The ARD measures the pairwise distances of probes within a contig and smoothes the distances of probes across different contigs. It shows distinct jumps at contig borders. This makes it appropriate for contig selection by clustering. We have designed a physical mapping algorithm that makes use of these observations and seems to be particularly well suited to the delineation of reliable contigs. We evaluated our method on data sets from two physical mapping projects. On data from the recently sequenced bacterium *Xylella fastidiosa*, the probe contig set produced by the new method was evaluated using the probe order derived from the sequence information. Our approach yielded a basically correct contig set. On this data we also compared our method to an approach which uses the number of supporting clones to determine contigs. Our map is much more accurate. In comparison to a physical map of *Pasteurella haemolytica* that was computed using simulated annealing, the newly computed map is considerably cleaner. The results of our method have already proven helpful for the design of experiments aimed at further improving the quality of a map.

Key words: clone-probe hybridization mapping, contig selection, bootstrap.

1. INTRODUCTION

The goal of physical mapping is to order a set of genetic landmarks or a library of cloned fragments of DNA according to their position in the genome. Physical maps are powerful tools for localization and isolation of genes, studying the organization and evolution of genomes and as a preparatory step for efficient sequencing. Even in the postgenome era, it is quite probable that genome-wide functional analyses will precede the sequencing of various organisms. For many such techniques, however, mapping information will still be an important requirement to see functions in their genomic perspective and also to make them accessible to function-directed sequence analysis. Different experimental techniques are used in physical mapping. Roughly, these are clone-probe hybridization mapping (Hoheisel *et al.*, 1993),

[1] German Cancer Research Center (DKFZ), Theoretical Bioinformatics (H0300), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

[2] German Cancer Research Center (DKFZ), Functional Genome Analysis (H0800), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

STS mapping (Hudson *et al.*, 1995), restriction mapping (Coulson *et al.*, 1995), radiation-hybrid mapping (Slonim *et al.*, 1997), and optical mapping (Lin *et al.*, 1999). Here we focus on a physical mapping strategy based on hybridization experiments (Hoheisel *et al.*, 1993; Scholler *et al.*, 1995; Hanke *et al.*, 1998). This procedure starts with a library of clones which correspond to subintervals of a larger contiguous piece of DNA G, all subintervals having the same size. Experimentally this can approximately be achieved by size selection methods described in Hoheisel *et al.* (1996).

In a more formal setting, from this clone library CL we select a subset $P \subset CL$ of probes. Each probe $p_i \in P$ is labeled and tested against the clone library. If a clone contains sufficient sequence similarity to the probe sequence, the probe will hybridize to this clone and a positive hybridization signal can be detected. The result of these experiments is a binary *clone/probe hybridization matrix* $A = (a_{ij})$ where

$$a_{ij} := \begin{cases} 1 & \text{if probe } p_j \text{ hybridizes to clone } c_i; \\ 0 & \text{otherwise.} \end{cases}$$

The physical mapping problem is to find the order of the probes in P that corresponds to their real position in G. A subsequent problem would then be to extend this order to the whole clone library. Here, we do not deal with the latter question, though. In the error-free case, the physical mapping problem can be translated into the following optimization problem (Greenberg and Istrail, 1995): Given a hybridization matrix, find a permutation of the columns (probes) such that the reordered matrix has the *consecutive ones property*, i.e., every row has at most one block of consecutive ones.

Unfortunately, physical mapping by hybridization experiments is highly influenced by errors and ambiguities: there are high rates of false positive and false negative hybridization signals and inconsistent hybridization signals caused by repetitive sequences, chimeric clones, or clones containing deletions. Additionally, there is variation in library coverage and in clone size. Note that even in the error-free case ambiguities may occur due to multiple solutions to the consecutive ones problem.

In the absence of errors, all admissible probe orders can be found and characterized efficiently using the *PQ-tree* data structure defined in Booth and Lueker (1976). However, in the presence of noise there is no generalization of the PQ-tree approach and the problem becomes ill defined. Our approach to this problem can be described as follows: we partition the probe set into independent parts (probe contigs). Based on these probe contigs, we clean the hybridization data. Then the probes are ordered inside the probe contigs. Finally the data is reinvestigated and additional experiments are suggested in order to improve and extend the map. This procedure can be iterated several times. In the rest of the paper we will focus only on the partitioning of the probe set into probe contigs, the essential step in the procedure.

Our mapping strategy is based on clustering of probes under a particular distance function. This distance is based on the evaluation of rank differences of probe orders as derived from multiple bootstrap replicates of the original hybridization data. We demonstrate certain properties of this distance function on idealized data that we believe make it particularly appropriate for use in conjunction with a clustering algorithm. The result of the clustering is a partitioning of probes into contigs. We also present methods to order the probes within the contigs.

There are several computational approaches which could be adapted for our physical mapping setting. Most of them globally optimize a certain objective function to construct a preliminary order for all markers/clones and offer then the possibility of interaction to improve this order. In the context of STS-content mapping, Alizadeh *et al.* (1995a,b) present both a detailed formal analysis and several computational approaches for finding a good marker order. This work contains an approach which relies on maximizing the posterior probability of a marker order, an approach which relies on solving the Hamming distance travelling salesman problem (TSP) and on algorithms for obtaining a good initial probe order and for data cleaning. The authors also discuss and evaluate several combinations of these methods. Cuticchia *et al.* (1992) use simulated annealing to order a clone set according to a binary clone fingerprint, implemented in the program ODS. Wang *et al.* (1993) use a random cost algorithm to order a clone set according to objective functions based on the Hamming distance of binary clone fingerprints. Mott *et al.* (1993) describe the programs PROBEORDER, BARR, and COSTIG which use simulated annealing and tree-search techniques to compute a map based on a maximum-likelihood distance measure between neighboring probes. SEGMAP (Green and Green, 1991) is a powerful interactive graphical tool for analyzing STS-content data which computes an optimal marker order by exhaustively rearranging some supplied suboptimal orders. In the special settings of unique end-probes and nonoverlapping probes, Christof *et al.* (1997) and Christof and Kececioglu (1999) apply a branch-and-cut approach to determine a probe order.

Computational approaches which primarily divide the data into different contigs before computing a marker order are the mapping software CONTIGMAKER of the WI/MIT group (Hudson *et al.*, 1995) and the program CONTIG EXPLORER described in Nadkarni *et al.* (1996). In contrast to our approach, they rely on the number of clones which share a certain probe pair for their contig definition.

Bootstrap resampling was introduced in Efron (1979) as a computer-based method for assigning measures of accuracy to statistical estimates. In physical mapping, Wang *et al.* (1994) and Liu (1998) used this technique to determine the reliability of a clone/marker order. A clustering strategy similar to ours was used in Mayraz and Shamir (1999) in the context of oligonucleotide fingerprinting. An introduction to rank correlation methods can be found in Kendall (1970).

The following section contains the basic definitions and algorithms. It starts by summarizing the initial steps of our procedure where we draw on established methods first for computing one physical map and then for bootstrapping. Next, the *averaged rank distance* on probes will be defined. Properties of this distance follow. The clustering algorithm presented afterwards uses this distance. In the Results section we apply our method to maps of *Xylella fastidiosa* and *Pasteurella haemolytica*. An assessment of the approach and some directions for future development are given in the Discussion section.

2. ALGORITHMS

Our strategy is the following. First we repeatedly apply a standard map construction algorithm based on simulated annealing to bootstrap resamplings of the hybridization data. The resulting bootstrap replicates form the basis for our probe distance function, the *averaged rank distance*. This distance is then used for constructing contigs by a modified clustering method. Finally, the probes within a contig need to be ordered.

2.1. Basic algorithm for map construction

We focus on ordering the probe set P. To compute the order of probes in P we use a vector-TSP (Cuticchia *et al.*, 1992; Alizadeh *et al.*, 1995a,b) formulation based on the Hamming distance between the columns of the clone/probe hybridization matrix A. The probe set P is extended by a dummy probe p_0 to yield $\widetilde{P} := P \cup \{p_0\}$ and likewise the hybridization matrix A is extended by a dummy column consisting only of 0's to give $\widetilde{A}$. We construct a complete weighted graph $G = (\widetilde{P}, E, c)$ where weight $c((p_i, p_j))$ is defined as the Hamming distance of column i and j in $\widetilde{A}$. Now the optimization problem consists of finding in G a Hamiltonian cycle of minimal weight. Such a minimal Hamiltonian cycle corresponds to a probe order which minimizes the number of blocks of consecutive ones in the hybridization matrix with reordered probes. This order is supposed to approximate the correct solution (Greenberg and Istrail, 1995; Xiong *et al.*, 1996). For the minimization we use the simulated annealing algorithm of Press *et al.* (1992).

2.2. Bootstrap resampling

In order to simulate independent replications of the physical mapping experiment *in silico* we resample the data set using a bootstrap strategy (Efron, 1979) which is similar to the approach of Wang *et al.* (1994); however, with the roles of clones and probes interchanged. We create new hybridization data matrices by resampling $|CL|$ times with replacement from the rows of A. This corresponds to repeating the hybridization experiments using the same set of probes P but creating new clone libraries by resampling from the original clone library CL.

In order to determine how often the procedure had to be reproduced, we tested the variance in independent experiment repetitions using different numbers of bootstrap replicates. With more than 200 resamplings, the results are well reproducible.

2.3. Averaged rank distance

While "contig" usually refers to an ordered set of overlapping clones representing a contiguous stretch of DNA, we here introduce the notion of a probe contig.

Let $P = \{p_1, \ldots, p_n\}$ denote the set of given probes and let Π be a family of permutations of P. Π may, for example, be the result of bootstrapping the physical mapping data. Then $C = \{p_{i_1}, \ldots, p_{i_m}\} \subset P$ is a *probe contig* if it is a maximal set of probes occurring as a "fixed block" in all permutations of Π. This means the probes occur continuously in a fixed order $\overrightarrow{C} = (p_{i_1}, \ldots, p_{i_m})$ or its reverse $\overleftarrow{C} = (p_{i_m}, \ldots, p_{i_1})$ in each permutation of Π and there is no superset of C with this property.

As an example, consider a set of probes $P = \{p_1, \cdots, p_5\}$ and a family of permutations $\Pi = \{(p_1, p_2, p_5, p_4, p_3), (p_2, p_1, p_3, p_4, p_5)\}$. This yields two probe contigs, $C_1 = \{p_1, p_2\}$ and $C_2 = \{p_3, p_4, p_5\}$.

In contrast to the idealized definition of probe contigs, when investigating bootstrap replicates of physical mapping experiments, one typically finds sets of probes where the interior order and integrity are only approximately maintained. This is due partly to the particular data selection in the bootstrap replicates, partly to suboptimal optimization in map construction (simulated annealing), and partly to ambiguity in the raw data. Therefore, in order to determine the probe contigs of a physical map by investigating bootstrap replicates, we use a distance function between probes that tries to correct for this fuzziness.

Let $\mathrm{rk}_\pi(p_i)$ denote the position (*rank*) of probe p_i in permutation π. Given a family of probe permutations Π, the *averaged rank distance* (ARD) between two probes p_i and p_j is defined as

$$ARD_\Pi(p_i, p_j) := \frac{1}{|\Pi|} \sum_{\pi \in \Pi} |\mathrm{rk}_\pi(p_i) - \mathrm{rk}_\pi(p_j)|.$$

We omit the subscript Π when there is no ambiguity. This distance averages the rank distances of probes in the bootstrap replicates. The idea is that, in the different bootstrap replicates, the probes which belong to the same contig should occur close to each other with a high reliability even if their correct order is not exactly defined. In contrast, the position and orientation of different contigs should be random and therefore the distances of probes belonging to different contigs should be significantly higher and show a higher variability. In the following, we show some properties of the ARD.

Theorem 1. *The averaged rank distance is a metric.*

Proof. The rank distance $d_\pi(p, q) := |\mathrm{rk}_\pi(p) - \mathrm{rk}_\pi(q)|$ of two elements $p, q \in P$ in a permutation $\pi \in \Pi$ is a metric. Therefore the average of these values over all permutations is a metric as well. ■

Theorem 2. *Within a probe contig $\overrightarrow{C} = (p_1, p_2, \ldots, p_m)$ the ARD distance between p_i and p_j is $|i - j|$ (see Figure 1).*

Proof. By the definition of probe contig this property holds for each permutation $\pi \in \Pi$, and hence it also holds for the average. ■

Our intention is to analyze the permutations resulting from bootstrapping a physical mapping experiment. In those permutations, we observed that while the contig structure is generally maintained, there seems to be no preference as to the order in which contigs occur. Likewise, there is no obvious preference as to the orientation of the individual contigs. To model this behavior, we define for a given set of contigs the space $\widetilde{\Pi}$, which consists of all possible probe permutations compatible with the contig set. More precisely, for each possible contig order and each contig occurring in its two orientations, $\widetilde{\Pi}$ contains the implied probe permutation.

⋮								
p_7	6	5	4	3	2	1	0	
p_6	5	4	3	2	1	0	1	
p_5	4	3	2	1	0	1	2	
p_4	3	2	1	0	1	2	3	
p_3	2	1	0	1	2	3	4	
p_2	1	0	1	2	3	4	5	
p_1	0	1	2	3	4	5	6	
	p_1	p_2	p_3	p_4	p_5	p_6	p_7	⋯

FIG. 1. The ARD distance matrix within a probe contig.

Theorem 3. *Let C_1, C_2 be two probe contigs, $C_1 \neq C_2$. Then for all $p_1, p_2 \in C_1, q_1, q_2 \in C_2$,*

$$ARD_{\widetilde{\Pi}}(p_1, q_1) = ARD_{\widetilde{\Pi}}(p_2, q_2) = const.$$

Proof. Let $\overrightarrow{C_1} = (p_1, \dots, p_k)$ and $\overrightarrow{C_2} = (q_1, \dots, q_l)$. We will show that

$$\underset{\widetilde{\Pi}}{ARD}(p_i, q_j) = \underset{\widetilde{\Pi}}{ARD}(p_{i-1}, q_j) \quad \text{for} \quad 1 < i \leq k. \tag{1}$$

First, let $C_1 <_\pi C_2$ if and only if $\mathrm{rk}_\pi(p) < \mathrm{rk}_\pi(q)$ for all $p \in C_1$ and $q \in C_2$. (Note that this is a valid definition by our definition of a probe contig.) Then, following immediately from the definition of the ARD and the property that all permutations of probe contigs occur equally often in $\widetilde{\Pi}$, we have

$$\underset{\widetilde{\Pi}}{ARD}(p_i, q_j) = \frac{1}{2}\left(\underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2\}}{ARD}(p_i, q_j) + \underset{\{\pi \in \widetilde{\Pi} : C_2 <_\pi C_1\}}{ARD}(p_i, q_j) \right).$$

Similarly, using the definition of the ARD and the property of $\widetilde{\Pi}$ that in $\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2\}$ both orientations of C_1 occur equally often, we have

$$\underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2\}}{ARD}(p_i, q_j) = \frac{1}{2}\left(\underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overrightarrow{C_1}\}}{ARD}(p_i, q_j) + \underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overleftarrow{C_1}\}}{ARD}(p_i, q_j) \right).$$

Using

$$\underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overrightarrow{C_1}\}}{ARD}(p_i, q_j) = \underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overrightarrow{C_1}\}}{ARD}(p_{i-1}, q_j) - 1,$$

and the symmetric equality for the second term, we get

$$\begin{aligned} \underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2\}}{ARD}(p_i, q_j) &= \frac{1}{2}\left(\underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overrightarrow{C_1}\}}{ARD}(p_{i-1}, q_j) - 1 + \underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2 \wedge \overleftarrow{C_1}\}}{ARD}(p_{i-1}, q_j) + 1 \right) \\ &= \underset{\{\pi \in \widetilde{\Pi} : C_1 <_\pi C_2\}}{ARD}(p_{i-1}, q_j). \end{aligned}$$

Similarly we obtain $ARD_{\{\pi \in \widetilde{\Pi} : C_2 <_\pi C_1\}}(p_{i-1}, q_j) = ARD_{\{\pi \in \widetilde{\Pi} : C_2 <_\pi C_1\}}(p_i, q_j)$, and Equation (1) follows. From this one can easily derive the proposition. ■

Theorem 4. *Based on Theorem 3 one may speak of $ARD_{\widetilde{\Pi}}(C_1, C_2)$. There holds*

$$\begin{aligned} ARD_{\widetilde{\Pi}}(C_1, C_2) &= \frac{2|P| + |C_1| + |C_2|}{6} \qquad (2) \\ &\geq \frac{|P| + 1}{3}. \qquad (3) \end{aligned}$$

Proof. Suppose C_1 and C_2 are the only probe contigs. It is easy to verify that

$$ARD_{\widetilde{\Pi}}(C_1, C_2) = \frac{|C_1| + |C_2|}{2}.$$

Now suppose we have $k > 2$ probe contigs $C_1, C_2, \dots, C_k$. By the properties of $\widetilde{\Pi}$, each of the probe contigs $C_i \in C_3, \dots, C_k$ has probability $\frac{1}{3}$ to occur *between* the probe contigs C_1 and C_2. Therefore its contribution to $ARD_{\widetilde{\Pi}}(C_1, C_2)$ is $\frac{1}{3}|C_i|$. Summing up, we get

$$\sum_{i=3}^{k} \frac{1}{3}|C_i| = \frac{1}{3}(|P| - |C_1| - |C_2|).$$

Additionally we have the contribution of C_1 and C_2 as calculated above. Together Equality (2) follows. Inequality (3) is obvious. ■

Note that we can also generalize Theorems 3 and 4 to the case where the order of probes within the probe contigs is not constant, if we assume that this order is independent of the (internal) order, position, or orientation of other probe contigs. It suffices to show that all averages of the rank differences of p_i and q_j over all permutations where p_i and q_j are at fixed positions in C_1 and C_2, are constant. But this follows with exactly the same proof as above.

The motivation for defining the ARD distance was the relation between the bootstrap results and the contig permutations. ARD distance matrices based on probe orders derived from bootstrap replications of real data will be shown in Figures 3, 6, and 7. Theorem 4 predicts strong 'jumps' of the distance values at contig borders, which we indeed observe on the real data as well. Moreover, by Theorem 3, the distances between probes from two contigs should be constant, yielding a "chess board pattern." This feature, too, can be recognized on the real data. Thus, the idealized properties derived for ARD on $\tilde{\Pi}$ seem to describe bootstrap data quite well.

2.4. The contig construction algorithm

Encouraged by the results presented above, we proceed to utilize the ARD for clustering on probes in order to define contigs as clusters. The algorithm is similar to the map construction algorithm described in Mayraz and Shamir (1999). It is a modification of a greedy clustering algorithm, where a special contig distance function is combined with a merge criterion that decides which growing contigs may be merged, based on an intercontig distance.

In order to prepare for the algorithm, we first define the contig distance function and the merge criterion.

Contig distance function. Given two ordered probe contigs, $\overrightarrow{C_1} = (p_1, \cdots, p_k)$ and $\overrightarrow{C_2} = (q_1, \cdots, q_l)$, with $p_i, q_j \in P$ and a family of probe permutations Π, we consider all four possible concatenations $\tilde{C} = \{\overrightarrow{C_1}\overrightarrow{C_2}, \overrightarrow{C_1}\overleftarrow{C_2}, \overleftarrow{C_1}\overrightarrow{C_2}, \overleftarrow{C_1}\overleftarrow{C_2}\}$ and compute

$$d(C_1, C_2) := \min_{C \in \tilde{C}} \left\{ \frac{1}{|C_1||C_2|} \sum_{p \in C_1, q \in C_2} (ARD_\Pi(p,q) - |\operatorname{rk}_C(p) - \operatorname{rk}_C(q)|)^2 \right\}. \tag{4}$$

This value defines the *contig distance* of C_1 and C_2.

The contig distance measures the mean square deviation of the ARD values from an ideal ARD distance matrix corresponding to the putative linear order C. A similar distance was discussed in Weeks and Lange (1987) in the context of linkage analysis.

The merge criterion. In order to prevent merging different probe contigs, we test if their measured ARD values could be better explained by a merged contig pair or by two unmerged probe contigs. In analogy to the contig distance d, we define for two probe contigs C_1 and C_2 the *intercontig distance*

$$d^*(C_1, C_2) := \frac{1}{|C_1||C_2|} \sum_{p \in C_1, q \in C_2} (ARD_\Pi(p,q) - \frac{2|P| + |C_1| + |C_2|}{6})^2. \tag{5}$$

This function measures the mean square deviation of the ARD values from the ARD values for two independent probe contigs C_1 and C_2 as predicted by Theorem 4. For each putative pair of probe contigs C_1 and C_2 to be merged, we compare this value to the contig distance $d(C_1, C_2)$ and allow merging only if $d(C_1, C_2) < d^*(C_1, C_2)$.

The algorithm. We now describe the algorithm that, from a set of probes, constructs a set of probe contigs. It consists of three steps:

1. Initialize the contig set such that each single probe corresponds to a contig, $C_i := \{p_i\}$.
   Initialize distance matrices $D[i, j] := d(C_i, C_j)$ and $D^*[i, j] := d^*(C_i, C_j)$ for all (i, j).

```
2. Repeat while further merges are possible:
   a. Search the contig distance matrix for the smallest distance $D[i_0, j_0]$.
   b. If the merge criterion is fulfilled, i.e., if $d(i_0, j_0) < d^*(i_0, j_0)$
      i. merge contigs $C_{i_0}$ and $C_{j_0}$;
      ii. update the distance matrices.
      otherwise
      i. set the contig distance to infinity: $D[i_0, j_0] := \infty$.
3. Output the contig set.
```

Whenever two probe contigs are merged in Step 2.a.i., the corresponding orientation $C \in \widetilde{C}$ that yielded the minimum in (4) is used.

Updating the distance matrices in Step 2.a.ii. is straightforward. One first removes the rows and columns of C_1 and C_2 in D and D^*, and then one inserts a new row and a new column for the merged contig C where the distances, as in the initialization, are computed using Equations (4) and (5), respectively.

2.5. Probe ordering within a contig

We have already obtained a linear order of the probes within a contig, which results from the orientation of the contigs at the merging step in the contig construction algorithm. However, the computation of the order was not the primary goal of the clustering algorithm, and hence more sophisticated re-orderings might yield better results. We present two alternative possibilities:

1. We assign each clone to a single contig using a maximum likelihood approach similar to the algorithm for fitting clones to a probe order described in Mott *et al.* (1993) and erase its hybridization signals in other probe contigs. Now the order of probes within a contig can be recomputed by any physical mapping algorithm (for example the basic algorithm for map construction described in the Algorithm section), using only the hybridization data of the clone set which was assigned to this contig.
2. We can also form a "consensus" of the bootstrap maps. We first delete in each bootstrap map the probes which do not belong to the investigated contig. Then, for each of these maps, we determine the orientation which best fits the probe order obtained by the contig construction algorithm. Using this orientation, we rank all probes. If we now order the probes corresponding to the sum of their alloted ranks in the different bootstrap maps, it can be shown (Kendall, 1970) that this order has the highest averaged Spearman rank correlation to all bootstrap replicates and can therefore be used as a "consensus order."

2.6. Analysis and implementation of the algorithms

Assume k permutations (the bootstrap replicates) of the n probes are given. Then a straightforward algorithm that computes the $n(n-1)/2$ ARD values between all probes runs in total time $O(n^2k)$ and uses $O(n^2)$ space. Using these precomputed values, it is easy to compute for a given pair (C_1, C_2) the two values $d(C_1, C_2)$ and $d^*(C_1, C_2)$ in time $O(|C_1||C_2|)$. In particular, the complete initialization of tables $D[i, j]$ and $D^*[i, j]$ in Step 1 of the contig construction algorithm takes $O(n^2k)$ time.

It can easily be seen that, using a priority queue storing of the distance table D, the greedy clustering of n elements can be computed in time $O(n^2 \log n + nt)$ where t is the time required to compute all distances of a newly created (merged) cluster C to the remaining clusters. In our case, t is $O(|C|n)$, which is bounded by $O(n^2)$. Moreover, in our modified greedy clustering algorithm, before merging we have to test if the merge criterion is fulfilled. Each such test can easily be done in constant time. Hence, the complete clustering (Step 2 of the contig construction algorithm) takes $O(n^3)$ time in the worst case.

The algorithms for map construction and bootstrapping were written in C++ in the LEDA 3.8 environment (Melhorn and Näher, 1999). For solving the vector-TSP, we adapted the simulated annealing routine of Press *et al.* (1992). Visualizations of the distance and variance matrices were done in MATLAB, visualizations of the clone/probe hybridization matrices were done using the program package *Programs for Analysing Hybridisation Data*, version 2 by R. Mott and A. Grigoriev and described in Mott *et al.* (1993).

The complete computation for the *Pasteurella haemolytica* data set (255 probes and 1025 clones), including the 200 bootstrap resamplings, took about 135 minutes on a SUN Ultra Enterprise 450 with 400 MHz. Note that, using the bootstrap approach, our method obviously was not designed to run as fast as possible, but rather to yield results of the highest possible quality.

3. RESULTS

3.1. Validation of the ARD and the clustering

In order to validate our algorithm, we tested it on the hybridization data of *Xylella fastidiosa*. This data set was created by Frohme *et al.* (unpublished) for the *Xylella fastidiosa* Genome Project.

During the development of our algorithm the sequence was unknown. While finishing, however, the sequence became available (Simpson *et al.*, unpublished) such that we are now able to obtain the exact position of 181 probes in the genome. A visualization of the hybridization data matrix using this "correct" probe order (corresponding to the sequence position) is shown in Figure 2.

We used the hybridization data to create 1000 resampled hybridization matrices, and then we computed the corresponding probe orders and their ARD values. Figure 3 shows a visualization of the ARD values (left) and the variances of the ARD values (right) using the "correct" probe order. Apart from a few outliers which are persistently misplaced by our map construction algorithm, the ARD values show the structure predicted by Theorems 2, 3, and 4. On the main diagonal, one finds blocks of small values which correspond to probe contigs (Theorem 2). These blocks show distinct "jumps" at the borders (Theorem 4). Moreover, by Theorem 3 the distances between probes from two contigs should be constant yielding a "chess board pattern." This feature, too, can be recognized on the real data. Additionally, the variances of the ARD values (Figure 3, right) also confirm our prediction that ARD values within a probe contig should show a small variance compared to the variances between probes of different contigs. Our contig construction algorithm applied to this data set yielded twenty probe contigs including three singletons (see Table 1).

The selected contigs correspond to the blocks on the main diagonal of the ARD distance matrix and the corresponding variances. We found six incorrectly placed probes in the contig set: probes 14, 30, 37, 70, 126, 159. A re-examination of these probes on the sequence level yielded that probes 14 and 37 overlap with large repeats. Clearly, these probes were placed at a wrong occurrence of these repeats in the genome. Probe 70 also overlaps with a repeated sequence but the other occurrence does not match the position of this probe as well. The remaining probes 30, 126 and 159 have a misleading hybridization pattern (strong

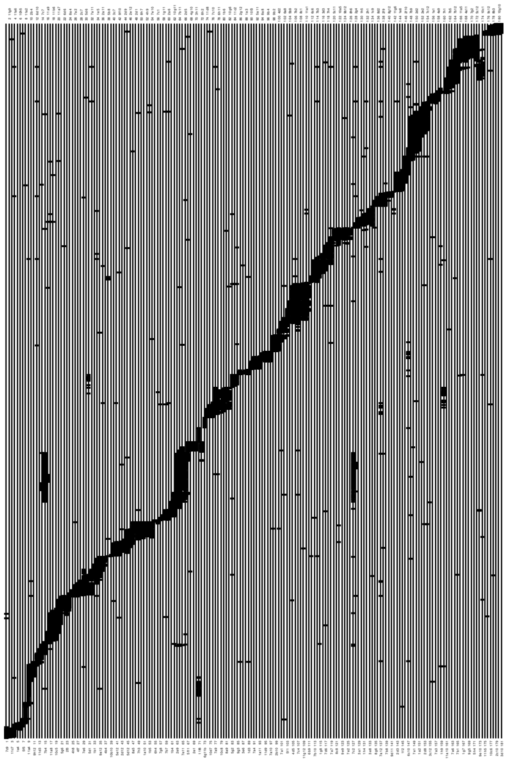

FIG. 2. A physical map of *Xylella fastidiosa* produced by procedures as described in Hoheisel *et al.* (1993) using the "correct" probe order.

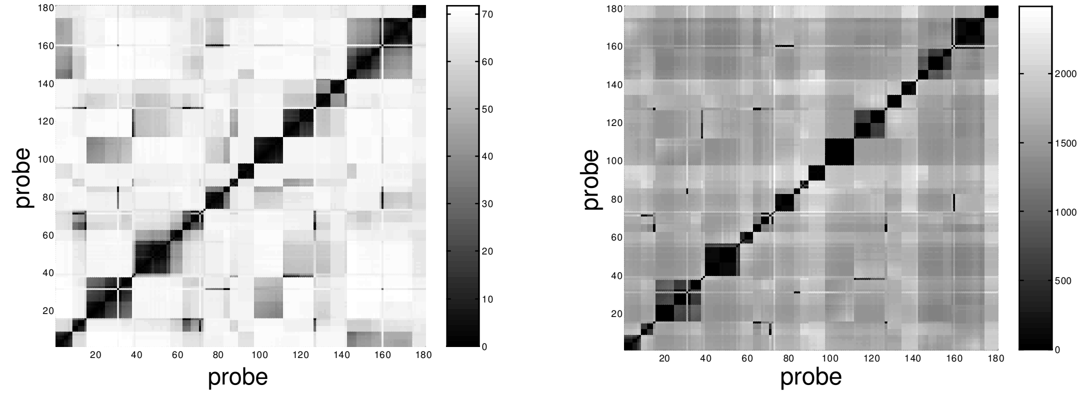

FIG. 3. ARD distance matrix of the *Xylella fastidiosa* data set using the "correct" probe order based on 1000 bootstrap replicates (**left**). Variance of these ARD values (**right**).

signal in another region of the genome) which cannot be explained on the sequence level and which we believe to result from a mix-up of clones.

Apart from these wrongly placed probes, the probe contigs are essentially correct, i.e., no contiguous stretches of probes of different positions in the genome are merged together in the same probe contig. It seems likely that repeat sequences may lead to increased ARD values by causing ambiguity in the probe order. In our case, this has not prevented the wrong placement of some single probes, but it prevented the algorithm from merging large probe stretches which do not belong together. It remains an interesting open

Table 1. Our Contig Construction Algorithm Applied to the *Xylella fastidiosa* Data Set Yielded 20 Probe Contigs (Left Column). Based on the Correct Order We Assigned Each Probe Contig a Position Corresponding to the Position of the Majority of its Probes (Center Column). Probes Inconsistent with This Position Were Counted as Wrongly Assigned and Are Listed in the Right Column ("−" Corresponds to Missing and "+" Corresponds to Wrongly Included Probes)

Contig	*Cosition*	*Wrongly assigned probes*
1	0–7	
2	8–13	+70
3	15–36	−30
4	38	
5	39–55	
6	56–61	
7	62–71	+14 +126 −70
8	72	
9	73–81	+159
10	82–84	+30
11	85–88	
12	89–96	
13	97–110	
14	111–125	+37
15	127–133	
16	134–140	
17	141	
18	142–157	
19	158–173	−159
20	174–180	

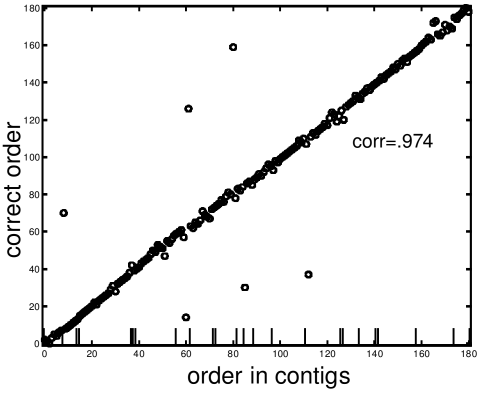

FIG. 4. Correlation of the probe order found by our clustering algorithm and the "correct" order. The contig borders are marked on the x-axis.

question to explore this behaviour in greater detail and to test the algorithm's performance on eukaryotic DNA whose repeats are much more complex than those of prokaryotes.

To demonstrate the quality of our clustering, we arranged the probe contigs in the correct order without changing the probe order inside the probe contigs (see Figure 4). Although our contig construction algorithm was mainly intended to compute a probe partition, it produced a remarkably good probe order.

Comparison with k-linked contigs. Hudson *et al.* (1995) (in the context of STS mapping) defined contigs based on the number of supporting clones. We adapted this method in the following way. Two probes are k-linked if they hybridize to at least k clones simultaneously. We assembled the *Xylella fastidiosa* probe set into k-linked contigs and evaluated the contig set. The result is shown in Table 2.

For values of k which produced a reasonable number of contigs, this approach always merges consecutive stretches of probes and results in a higher number of wrongly assigned probes compared to the results of our algorithm. We are well aware that the k-linkage approach is not designed to be a stand-alone method and that the resulting k-linked contigs could be improved by additional cleaning steps. Nevertheless, this demonstrates that the probe contig set, at least for our data sets, is a reasonable alternative to this approach.

3.2. Application to the Pasteurella haemolytica *data set*

In order to demonstrate the robustness of our clustering method we applied it to a noisy data set of *Pasteurella haemolytica* which is very difficult to process (Hanke *et al.*, unpublished). A conventional approach using simulated annealing to optimize the above-described vector-TSP formulation produced only an unsatisfactory result (Figure 5, left). A visualization of the ARD distance matrix (Figure 6, left) ordered with respect to this solution immediately highlights large regions which seem to be incorrectly ordered. A closer look with a higher magnification (Figure 7) also reveals local disorder.

Our cluster algorithm determined 39 contigs (Figure 6, right) which appear more homogeneous than the result derived by simulated annealing. We arranged these contigs (for presentation) in an order which minimizes the contig distance function (Equation 4). A visualization of the clone/probe hybridization matrix corresponding to this order is shown in Figure 5 (right). The improvements over the physical map based on simulated annealing (Figure 5, left) are obvious.

TABLE 2. THE NUMBER OF k-LINKED CONTIGS, MERGED CONTIGS, AND WRONGLY ASSIGNED PROBES AS COMPARED TO THE CORRECT PROBE ORDER OF THE *Xylella fastidiosa* DATA FOR DIFFERENT VALUES OF k

k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	···	*15*
Number of contigs	2	4	7	13	16	22	29	36	40	45	···	73
Merged contigs	—	2	2	5	5	7	5	3	3	3	···	2
Wrongly assigned probes	—	85	84	69	54	44	33	21	21	20	···	13

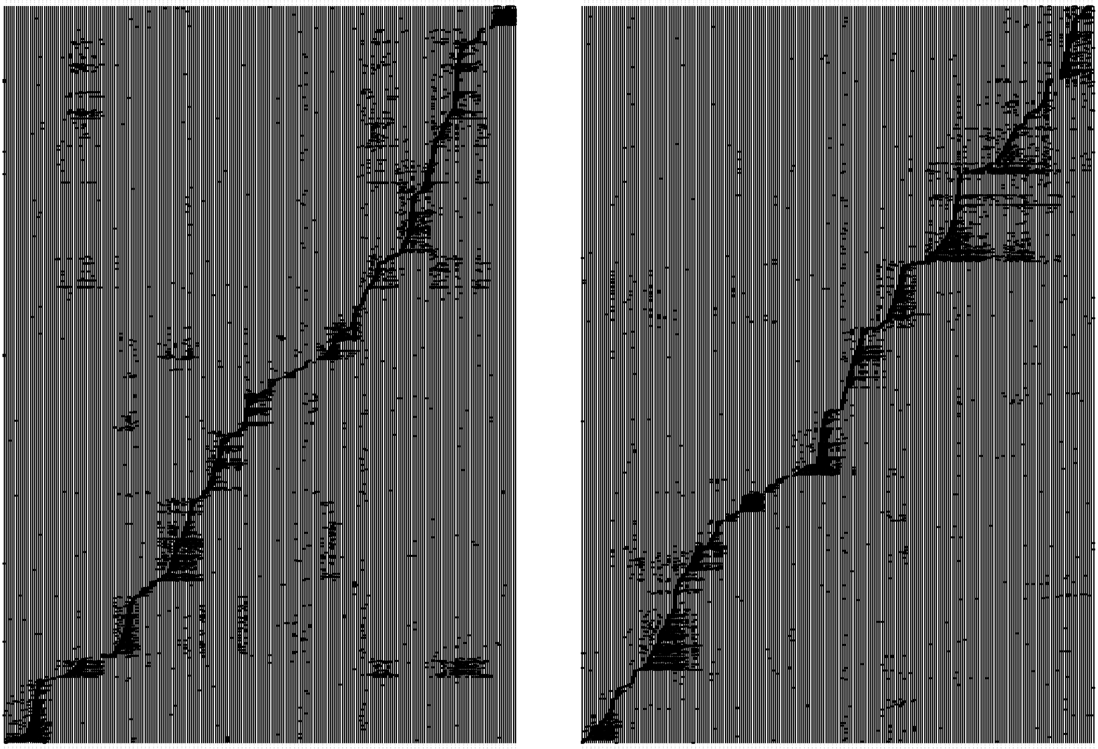

FIG. 5. Clone/probe hybridization matrix of *Pasteurella haemolytica* based on the best output of 200 simulated annealing runs (**left**). Map based on our cluster construction algorithm (**right**).

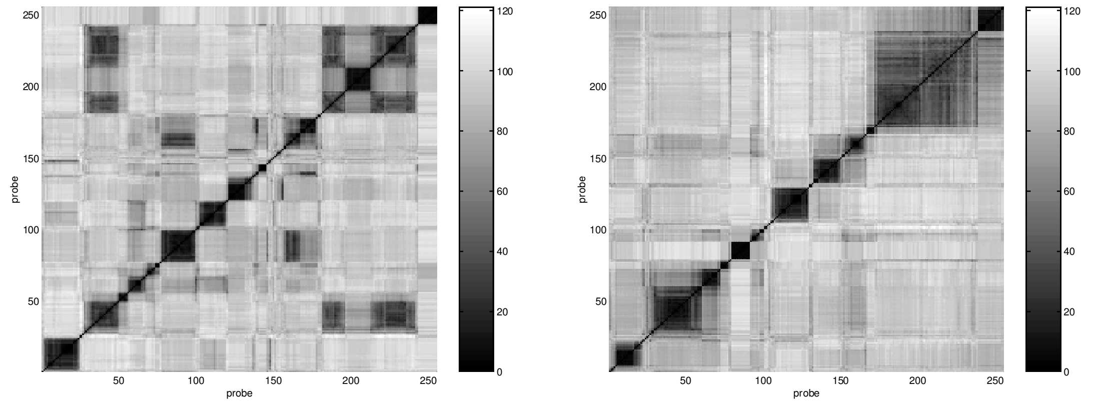

FIG. 6. ARD distance matrix of *Pasteurella haemolytica* ordered according to the simulated annealing result described in the text. The region p180–p240 shows putative disorder (left). The reordered ARD distance matrix is shown on the right.

p129	3.86	4.25	4.33	6.77	6.10	2.80	0.00
p128	5.22	5.07	4.50	6.39	5.19	0.00	2.80
p127	7.02	6.91	5.76	2.03	0.00	5.19	6.10
p126	6.79	6.45	5.75	0.00	2.03	6.39	6.77
p125	2.80	2.34	0.00	5.75	5.76	4.50	4.33
p124	1.20	0.00	2.34	6.45	6.91	5.07	4.25
p123	0.00	1.20	2.80	6.79	7.02	5.22	3.86
	p123	p124	p125	p126	p127	p128	p129

p127	7.02	6.91	5.76	6.10	5.19	2.03	0.00
p126	6.79	6.45	5.75	6.77	6.39	0.00	2.03
p128	5.22	5.07	4.50	2.80	0.00	6.39	5.19
p129	3.86	4.25	4.33	0.00	2.80	6.77	6.10
p125	2.80	2.34	0.00	4.33	4.50	5.75	5.76
p124	1.20	0.00	2.34	4.25	5.07	6.45	6.91
p123	0.00	1.20	2.80	3.86	5.22	6.79	7.02
	p123	p124	p125	p129	p128	p126	p127

FIG. 7. Enlargement of the ARD distance matrix of *Pasteurella haemolytica* ordered according to the simulated annealing result (**left**). We suppose that the probe order is locally incorrect. A configuration which fits better to an ARD distance matrix within a probe contig could be achieved if probes 129 and 128 (in this order) would be placed between probes 125 and 126 (**right**).

The results of our cluster algorithm have already been used in the *Pasteurella haemolytica* physical mapping project. At probe contig borders, additional probes for contig extension and gap closure were selected and used for additional hybridization experiments. Additionally, the computed probe order was used to select a clone set for ordering a plasmid library.

4. DISCUSSION

Clustering the set of probes into independent contigs and subsequently ordering these contigs is a natural approach to physical mapping. It divides the optimization problem into smaller and, it is hoped, easier subproblems that can be dealt with independently. At the same time, though, the danger is introduced of having errors in the contig selection which then propagate. In this work we presented a method for contig selection that apparently performs very well on real data.

The source of the robustness of the resulting contig definitions probably is twofold. First, bootstrapping is the *in silico* equivalent of repeating an experiment. For each resampled data set, we compute a physical map using a standard algorithm. Particularities of any one solution are lost and thus the sensitivity to outliers or peculiarities of the data is reduced.

Second, in order to combine the results of these computations we define a distance function between probes which averages the rank differences of probe pairs in these bootstrap maps. This approach can be interpreted as a generalization of the bootstrap procedure for physical mapping (Efron, 1979; Wang *et al.*, 1994; Liu, 1998) which not only takes into account the *consecutive* occurrence of two probes, but also uses the information of more distant connections. This leads to a robust and reliable distance function with interesting and useful properties. Averaged rank distance is largely independent of factors like coverage depth because it accounts only for distances in rank of probes. Within contigs, the averaged rank distance behaves much like other distance measures. Between contigs, however, individual distances between probes are less important because all probes in one contig tend to have roughly the same distance to any probe in a particular other contig. This between-contig distance tends to be much larger than the distances between neighboring probes in the same contig. In this respect, ARD on idealized data resembles an ultrametric in that all distances between elements of two clusters are equal. Hence, such a distance should be more easily approximated by a tree and allow for good clustering results.

The results shown are very encouraging. In addition, the distance matrices can also be used to visualize the reliability of a given probe ordering and to highlight dubious regions (see Figures 6, left, and 7). This has been shown very helpful to derive hypotheses about possible orderings and experiments which increase the quality of the map. Similar drawings for the bootstrap values are less meaningful because they incorporate only next neighbor connections.

Several lines of future work can be anticipated. The problem of contig construction is particularly challenging in physical mapping using STS-content data. For example, large STS mapping data sets were collected by the CEPH/Généthon and WI/MIT teams, but an assembly into comprehensive contig maps was impossible (Harley *et al.*, 1999). We plan to adapt our method to STS-content data and make it applicable to this kind of data. On the theoretical side, we are working on a probabilistic model that allows one to formulate the partitioning of probes into contigs as an optimization problem. Another interesting project could be to investigate the influence of other perturbation strategies, like subsampling, oversampling, or data perturbation, on our method.

ACKNOWLEDGMENTS

We are very grateful to the Organization for Nucleotide Sequencing and Analysis (ONSA), especially to João Carlos Setubal and João Meidauis, for providing us with the sequence data of the *Xylella fastidiosa* clone library. We would also like to thank Richard Desper for many helpful discussions. Jens Hanke provided us with the data set and physical map of *Pasteurella haemolytica*.

REFERENCES

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

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

Booth, K., and Lueker, G. 1976. Testing for the consecutive ones property, interval graphs and graph planarity using *PQ*-tree algorithms. *J. Comput. Syst. Sci.* 13, 333–379.

Christof, 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.

Christof, T., and Kececioglu, J. 1999. Computing physical maps of chromosomes with nonoverlapping probes by branch-and-cut, 115–123. In Istrail, S., Pevzner, P., and Watermann, M., eds., *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, ACM, New York.

Coulson, A., Huynh, C., Kozono, Y., and Shownkeen, R. 1995. The physical map of the Caenorhabditis elegans genome. *Methods Cell Biol.* 48, 533–550.

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

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 7, 1–26.

Green, E., and Green, P. 1991. Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods Appl.* 1, 77–90.

Greenberg, D., and Istrail, S. 1995. The chimeric mapping problem: Algorithmic strategies and performance evaluation on synthetic genomic data. *J. Comp. Biol.* 2, 219–274.

Hanke, J., Frohme, M., Laurent, J.-P., Swindle, J., and Hoheisel, J. 1998. Hybridization mapping of *Trypanosoma cruzi* chromosome III and IV. *Electrophoresis* 19, 482–485.

Harley, E., Bonner, A., and Goodman, N. 1999. Revealing hidden interval graph structure in STS-content data. *Bioinformatics* 15, 278–285.

Hoheisel, J., Maier, E., Mott, R., and Lehrach, H. 1996. Integrated genome mapping by hybridization, 319–346. In Birren, B., and Lai, E., eds., *Analysis of Non-Mammalian Genomes—A Practical Guide*. Academic Press, San Diego.

Hoheisel, J., Maier, E., Mott, R., McCarthy, L., Grigoriev, A., Schalkwyk, L., Nizetic, D., Francis, F., and Lehrach, H. 1993. High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. *Cell* 73, 109–120.

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

Kendall, M. 1970. *Rank Correlation Methods*. Griffin, London.

Lin, J., Qi, R., Aston, C., Jing, J., Anantharaman, T., Mishra, B., White, O., Daly, M., Minton, K., Venter, C., and Schwartz, D. 1999. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285, 1558–1562.

Liu, B. 1998. *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. CRC Press LLC, Boca Raton, FL.

Mayraz, G., and Shamir, R. 1999. Construction of physical maps from olignucleotide fingerprints data, 268–277. In Istrail, S., Pevzner, P., and Watermann, M., eds., *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, ACM, New York.

Melhorn, K., and Näher, S. 1999. *LEDA: A Platform for Combinatorial and Geometric Computing*. Cambridge University Press, Cambridge.

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*. *Nucl. Acids Res.* 21, 1965–1974.

Nadkarni, P., Banks, A., Montgomery, K., LeBlanc-Stracewski, J., Miller, P., and Krauter, K. 1996. CONTIG EXPLORER: Interactive marker-content map assembly. *Genomics* 31, 301–310.

Press, W., Teukolsky, W., Vetterling, W., and Flannery, B. 1992. *Numerical Recipes in C*. Cambridge University Press, New York.

Scholler, P., Karger, A., Meier-Ewert, S., Lehrach, H., Delius, H., and Hoheisel, J. 1995. Fine-mapping of shotgun template-libraries: An efficient strategy for the systematic sequencing of genomic DNA. *Nucl. Acids Res.* 23, 3842–3849.

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

Wang, Y., Prade, R., Griffith, J., Timberlake, W., and Arnold, J. 1993. A fast random cost algorithm for physical mapping. *Proc. Natl. Acad. Sci. USA* 91, 11094–11098.

Wang, Y., Prade, R., Griffith, J., Timberlake, W., and Arnold, J. 1994. ODS_BOOTSTRAP: Assessing the statistical reliability of physical maps by bootstrap resampling. *CABIOS* 10, 625–634.

Weeks, D., and Lange, K. 1987. Preliminary ranking procedures for multilocus ordering. *Genomics* 1, 236–242.

Xiong, M., Chen, R., Prade, R., Wang, J., Griffith, W., Timberlake, W., and Arnold, J. 1996. On the consistency of a physical mapping method to reconstruct a chromosome *in vitro*. *Genetics* 142, 267–284.

Address correspondence to:
Steffen Heber
German Cancer Research Center (DKFZ)
Theoretical Bioinformatics (H0300)
Im Neuenheimer Feld 280
D-69120 Heidelberg, Germany

E-mail: s.heber@dkfz-heidelberg.de