Genomics **69,** 235–241 (2000)
doi:10.1006/geno.2000.6325, available online at http://www.idealibrary.com on IDEAL®

Application of Bootstrap Techniques to Physical Mapping

Steffen Heber,*,†,[1] Jörg Hoheisel,† and Martin Vingron*

*Theoretical Bioinformatics and †Functional Genome Analysis, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received December 20, 1999; accepted July 13, 2000

Ordering genetic markers or clones from a genomic library into a physical map is a central problem in genetics. In the presence of errors, there is no efficient algorithm known that solves this problem. Based on a standard heuristic algorithm for it, we present a method to construct a confidence neighborhood for a computed solution. We compute a confidence value for putative local solutions derived from bootstrap replicates of the original solution. In the reliable parts, the confidence neighborhood and the computed solution tend to coincide. In regions that are ill-defined by the data, the neighborhood contains additional reasonable alternatives. This offers the possibility of designing further experiments for the badly defined regions to improve the quality of the physical map. We analyze our approach by a simulation study and by application to a dataset of the genome of the bacterium *Xylella fastidiosa*.

INTRODUCTION

The goal of physical mapping is to order a set of genetic markers 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, for studying the organization and evolution of genomes, and as a preparatory step for efficient sequencing. There are a wide variety of experimental techniques for physical mapping. The leading methods are clone–probe hybridization mapping (Hoheisel *et al.*, 1993), 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 protocol based on hybridization experiments (Hoheisel *et al.*, 1993; Scholler *et al.*, 1995; Hanke *et al.*, 1998).

The procedure can be described as follows. We start with a clone library C of clones that correspond to subintervals of a larger contiguous piece of DNA G, all having approximately the same size. From C we select a subset $P \subset C$ of probes P. Each probe $p_i \in P$ is labeled and tested against the clone library. If a clone contains DNA complementary 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 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. The physical mapping problem can be translated into the following combinatorial 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 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 by 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. The major practical problems in hybridization mapping are the management and visualization of large datasets, the efficient selection of probes to minimize the number of hybridization exper-

[1] To whom correspondence should be addressed at Funktionelle Genomanalyse (H0800)/Theoretische Bioinformatik (H0300), Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. Telephone: +49+6221+42 2720. Fax: +49+6221+42 2849. E-mail: s.heber@dkfz.de.

0888-7543/00 $35.00

iments, and the detection and resolution of inconsistencies in the hybridization data.

There are several computational approaches of STS-content map assembly that could be used for our protocol. Mott *et al.* (1993) developed 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. CONTIGMAKER was developed by the WI/MIT group (Daly *et al.*, 1994). The program clusters markers into double-linked contigs. These contigs are subsequently ordered using genetic and radiation-hybrid data. ODS is a program designed by Cuticchia *et al.* (1992) using simulated annealing to order a clone set according to a binary clone fingerprint. CONTIG EXPLORER is a program for exploratory sensitivity analysis and interactive map assembly (Nadkarni *et al.*, 1996). SEGMAP (Green and Green, 1991) is an interactive graphical tool for analyzing STS-content data. It computes an optimal marker order by exhaustively rearranging some supplied suboptimal orders. These program packages typically construct a preliminary probe order that optimizes a special objective function and offer the possibility of interaction to improve this order.

Additional desirable features of a physical mapping algorithm are (according to Setubal and Meidanis, 1997, p. 152) that it should distinguish "good" parts of the solution from "not so good" parts and that if several candidate solutions meet the optimization criteria, all of them should be reported. This could greatly facilitate further experiments.

In our attempt to add these features to the existing algorithms we assess the reliability of putative probe configurations. We use a bootstrap approach for this purpose. Bootstrap resampling was introduced by Efron (1979) as a computer-based method for assessing measures of accuracy to statistical estimates. In bioinformatics it is used in phylogeny (for an introduction, see Felsenstein, 1985) and also in linkage analysis (Liu, 1998). In physical mapping, Wang *et al.* (1994) used this technique to determine the reliability of a clonal ordering. Here we present a strategy that relies on the solution of a conventional physical mapping algorithm but extends this approach by creating a suitable neighborhood of this solution.

In principle, our strategy will work with every physical mapping algorithm that produces as output a single probe order and that is fast enough to be repeated several times. For concreteness we focus on an algorithm that uses a vector–TSP approach (Cuticchia *et al.*, 1992; Alizadeh *et al.*, 1995). We resample the *clone library* and create bootstrap replicates π^{*b}, $b = 1, \ldots, B$ of the original solution π. For each probe pair (p, q) with $p \neq q$ we define the bootstrap value $b((p, q)) \in [0, 1]$ as the frequency of the *consecutive* occurrence of p and q in the bootstrap replicates. We represent these values in the *bootstrap graph GB*, a graph on the probe set with the bootstrap values as weights. The true probe order corresponds to a Hamiltonian path in GB enumerating the probes in the order of their occurrence in the genome.

Based on the bootstrap graph GB we define a confidence neighborhood N of π. In the parts of the probe order π that are well supported by the data the corresponding bootstrap values are high and N does not differ from π. In regions that are not so good the bootstrap values are small and N contains additionally several other "reasonable" probe configurations. Using these configurations one can derive for the "bad" regions alternative probe orders that are useful for the design of additional experiments.

The rest of the paper is organized as follows. Under Materials and Methods we explain the different components of our approach. We describe the algorithm for map construction and the bootstrap strategy. Then we define the confidence neighborhood. Under Results we present a simulation study in which we determine the necessary number of bootstrap replicates and evaluate our approach. We also apply our method to the data set of the bacterial genome of *Xylella fastidiosa*. Under Discussion we give an assessment of the approach and some directions for future development.

MATERIALS AND METHODS

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 formulation (Cuticchia *et al.*, 1992; Alizadeh *et al.*, 1995) 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 $\tilde{P} := P \cup \{p_0\}$ and likewise the hybridization matrix A is extended by a dummy column consisting only of zeros to give $\tilde{A}$. We construct a complete weighted graph $GH = (\tilde{P}, E, c)$ in which weight $c((p_i, p_j))$ is defined as the Hamming distance of columns i and j in $\tilde{A}$. Now the optimization problem consists of finding a Hamiltonian cycle of minimal weight in GH. Such a minimal Hamiltonian cycle corresponds to a probe order that minimizes the number of blocks of consecutive ones in the hybridization matrix with reordered probes. This order is supposed to approximate the true solution (Greenberg and Istrail, 1995; Xiong *et al.*, 1996). For the minimization we use the simulated annealing algorithm of Press *et al.* (1992).

Bootstrap resampling. To simulate independent replications of the physical mapping experiment *in silico*, we resampled the dataset, using a bootstrap strategy similar to the approach of Wang *et al.* (1994), but with the roles of clones and probes interchanged. We created a new hybridization data matrix by selecting $|C|$ times with replacement from the rows of A. This corresponds to repeating the hybridization experiments using the same set of probes P, but creating a new clone library by resampling from the original clone library C. This procedure was repeated B times to obtain B resampled datasets.

For each of these resampled data sets, we computed a corresponding probe order π^{*b} using the above described simulated annealing approach (Basic algorithm for map construction). Let Π^* be the set of these B permutations. For each pair of probes (p, q) with $p, q \in \tilde{P}$ and $p \neq q$ we define the bootstrap value $b((p, q))$ as the relative frequency of their consecutive occurrence in Π^*, i.e.,

$$b((p, q)) := \frac{|\{\pi^* \in \Pi^* : |rk_{\pi^*}(p) - rk_{\pi^*}(q)| = 1\}|}{B},$$

for $p, q \in \tilde{P}$ and $p \neq q$.

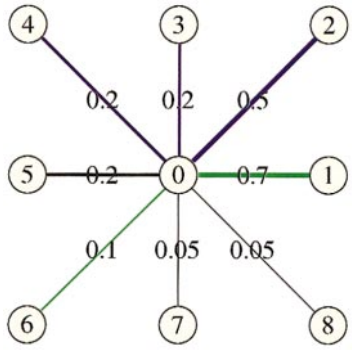


FIG. 1. Visualization of CI(0, 0.8). The edges adjacent to node 0 are sorted according to their bootstrap values. The edges that belong to $E(\pi)$ are colored green. CI(0, 0.8) consists of all edges that are colored blue or green.

Here $rk_{\pi^*}(p)$ is the rank of p in π^*. One can easily see that $b((p, q)) = b((q, p))$ and $b((p, q)) \in [0, 1]$. Using the fact that each $p \in \tilde{P}$ has exactly two different neighbors in each $\pi^* \in \Pi^*$, we obtain for each $p \in \tilde{P}$

$$\sum_{q \in \tilde{P}} b((p, q)) = 2. \quad [1]$$

The above-defined bootstrap values are represented in a weighted, complete bootstrap graph GB = $(\tilde{P}, E, b)$. In GB a probe order π corresponds to a Hamiltonian cycle—in the following we use $E(\pi) \subseteq E$ to indicate the corresponding edge set.

Confidence neighborhood. To represent the variability of a solution π of the basic algorithm for map construction, we define a neighborhood N of π in the bootstrap graph GB. Given π, the bootstrap graph GB = $(\tilde{P}, E, b)$, and a confidence level $\gamma \in [0, 1]$, we construct for each probe $p \in \tilde{P}$ a set of adjacent edges $CI(p, \gamma) \subseteq E$. To compute $CI(p, \gamma)$, the edges adjacent to the node p are sorted. First the edges $e_1, e_2 \in E(\pi)$ of the original solution π are included. Then, additional edges are taken into $CI(p, \gamma)$, heavier edges before lighter ones, until the summed bootstrap values of the edges in $CI(p, \gamma)$ exceeds 2γ (see Fig. 1 for a visualization). Edges with equal bootstrap values may occur during this procedure. Although the resolution of draws could be based on further analyses of the data, we found a random selection sufficient.

By construction, $CI(p, \gamma)$ is a minimal set of edges adjacent to p that contains the adjacent edges of $E(\pi)$ and fulfills the condition $\sum_{e \in CI(p,\gamma)} b(e) \geq 2\gamma$. The motivation is that, if one assumes that the bootstrap value of an edge adjacent to p corresponds to its "probability" of being part of the true solution, and that edges are independent, then for $\gamma \in (0.5, 1]$ we can interpret $CI(p, \gamma)$ as a $100 \cdot (2\gamma - 1)\%$ confidence interval for the true edges adjacent to p.

We define $N(\gamma) := \cup_{p \in \tilde{P}} CI(p, \gamma)$. By definition,

$$N(\gamma_1) \subseteq N(\gamma_2) \text{ for } \gamma_1 \leq \gamma_2 \text{ and } \gamma_i \in [0, 1]. \quad [2]$$

Thus we have a monotonically increasing parameterized candidate set for the true solution. For $\gamma = 0$ the set $N(0) = E(\pi)$ corresponds to the original solution π, while for $\gamma = 1$ we have $N(1) = E(\pi) \cup \{e \in E : b(e) > 0\}$. A simulation study (see Confidence Neighborhood) shows the relation between the size of N and the number of true edges included in N with respect to γ. A visualization of $N(0.95)$ for the dataset of the bacterial genome of *X. fastidiosa* is shown in Figs. 6 and 7.

Computation. The algorithms for map construction, map visualization, and bootstrapping were written in C++ in the LEDA 3.8 environment (Melhorn and Näher, 1995). To solve the vector–TSP, we adapted the simulated annealing routine of Press *et al.* (1992). Diagrams were done in MATLAB. The complete computation time for a dataset consisting of 200 probes and 1000 clones and a resampling rate of $B = 1000$ on a SUN Ultra Enterprise 450 with 400 MHz was approximately 13.5 h.

RESULTS

To test the behavior of our algorithms, we performed a simulation study. We created 50 artificial raw datasets mimicking the parameters of previous mapping projects (Hoheisel *et al.*, 1993; Scholler *et al.*, 1995; Hanke *et al.*, 1998). A linearized genome G of size $|G| = 2000$ kb was used to create a clone library of 1000 clones of equal size, $l = 40$ kb. This represents a 20-fold clone coverage. Clone start points were chosen uniformly from the interval $[1, |G| - l + 1]$. We selected 200 clones of the clone library and used them as probes corresponding to a 4-fold probe coverage. For each probe, a virtual hybridization experiment was simulated. A hybridization signal of a clone in the clone library was detected if the probe and the clone showed an overlap of more than 2 kb. Additionally, we added to the hybridization data false positive signals at a rate of 1% and false negatives at a rate of 5.5%.

How Many Bootstrap Replications Are Necessary?

There is no general rule to determine how many bootstrap replications B are necessary. For confidence intervals normally 1000–2000 are recommended (Efron and Tibshirani, 1993). Our goal was to ensure that the computed bootstrap values would be reproducible upon independent experiment repetitions. Therefore, we chose one of the artificially created datasets and computed 10 times independently the bootstrap values, using B replicates each time. For each possible pair of these 10 different experiments, we compared *all* corresponding bootstrap values and computed their maximal absolute difference.

Figure 2 shows a box plot of these differences for different resampling rates B. We chose $B = 1000$ resamplings for further experiments. This choice represents a compromise between our goals to guarantee reproducibility and to limit computation time.

Correlation of Bootstrap Values and Error Rate

To investigate the relation between the bootstrap value and the consecutive occurrence of probes in the true physical map, we evaluated the simulated datasets. We computed the corresponding bootstrap graphs with a resampling rate of $B = 1000$. Their edges were partitioned into 11 bins according to their bootstrap value. An edge $e = (p, q)$ was classified as "true" if p and q occurred consecutively in the "true" physical map and "false" otherwise (Table 1).

When averaged over 50 independent simulations, the vast majority of the false edges (19,351.98 or 96% of the false edges) are in the bin with bootstrap value 0. In the other bins, the number of false edges decreases monotonically as the bootstrap value increases. On the

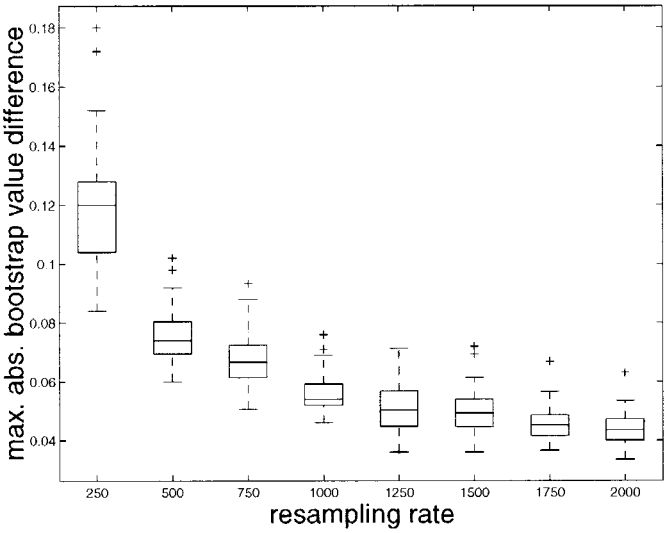


FIG. 2. Box plot of the maximal absolute bootstrap differences for different resampling rates. The lower and upper lines of the "boxes" are the 25th and 75th percentiles of the sample. The line in the middle is the sample median. The "whiskers" show the extent of the sample unless there are outliers (marked by +).

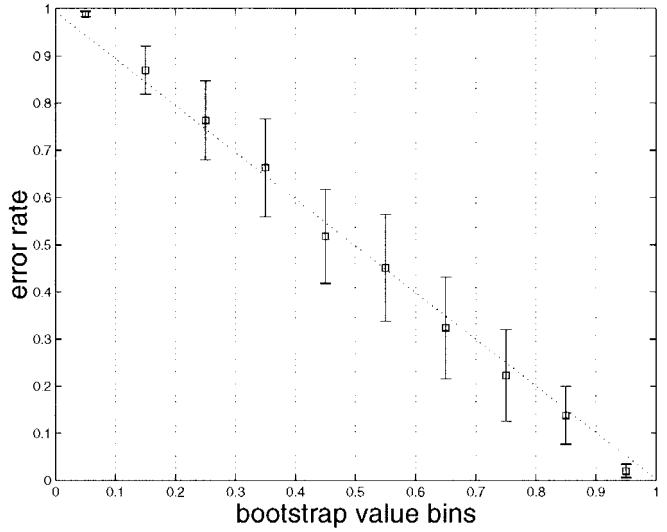


FIG. 3. Mean and standard deviation of the error rate in the different bootstrap value bins. The bin consisting of edges with bootstrap value 0 is omitted.

other hand, most of the true edges are in the high-scoring bootstrap bins (162.88 or 81% of the true edges have bootstrap values larger than 0.5), while only 0.06 (0.03%) true edges are in the bin with bootstrap value 0. This shows that the edges with bootstrap values strictly greater than 0 are suitable candidates for the true solution of the physical mapping problem.

If one defines the *error rate* within a bootstrap value bin as

$$\text{error rate} := \frac{|\{\text{false edges in bootstrap value bin}\}|}{|\{\text{bootstrap value bin}\}|},$$

then there is a strong negative correlation between error rate and bootstrap value (Fig. 3). Only minor deviations can be seen at the left and right margins. This is the motivation for us to use bootstrap values as a measure of "quality" for the physical map. For comparison: there was a mean error rate of 0.1694 (standard deviation 0.0324) in the edge set $E(\pi)$ corresponding to the result of our basic algorithm for map construction.

TABLE 1

Average Number of True and False Edges in the Bootstrap Bins

Bootstrap bin	True	False
$b = 0$	0.06	19,351.98
$0.0 < b \leq 0.1$	5.02	429.72
$0.1 < b \leq 0.2$	6.24	40.92
$0.2 < b \leq 0.3$	7.00	22.70
$0.3 < b \leq 0.4$	7.32	14.34
$0.4 < b \leq 0.5$	12.48	13.50
$0.5 < b \leq 0.6$	11.38	9.40
$0.6 < b \leq 0.7$	12.24	5.68
$0.7 < b \leq 0.8$	16.80	4.86
$0.8 < b \leq 0.9$	24.72	3.94
$0.9 < b \leq 1.0$	97.74	1.96

Confidence Neighborhood

To investigate the relation among the confidence level γ, the size of the confidence neighborhood $N(\gamma)$, and the number of true edges contained in $N(\gamma)$, we further evaluated the simulated datasets. Figure 4 is a plot of the number of true edges contained in $N(\gamma)$ versus the size of $N(\gamma)$. Intuitively, this indicates the price, measured in false edges, one has to pay for the delineation of an increasing number of true edges. When inspecting only a small neighborhood (γ small) one identifies a restricted number of mostly true edges. Upon increasing γ the size of the confidence neighbor-

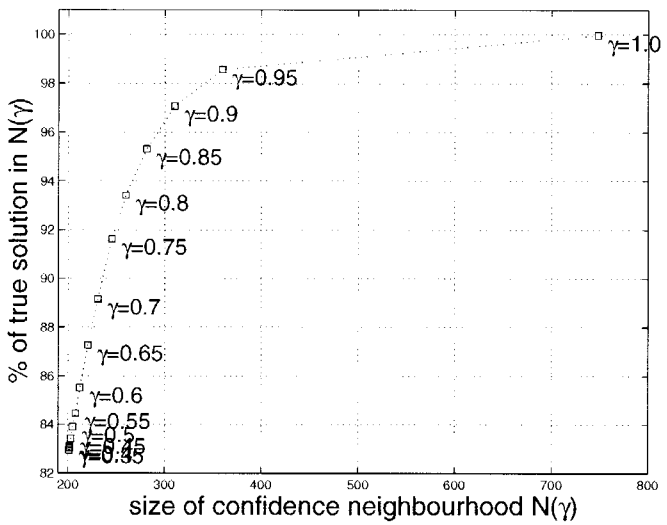


FIG. 4. Average number of true edges in the confidence neighborhood $N(\gamma)$ plotted against the size of $N(\gamma)$. The data points are labeled by their confidence level γ.

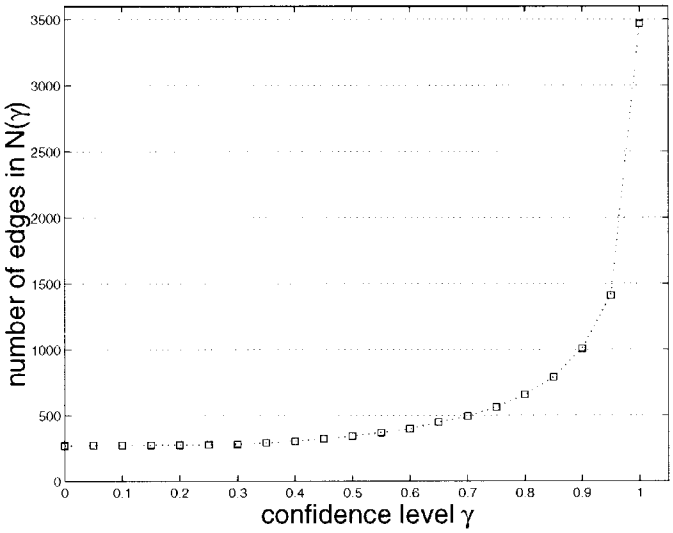


FIG. 5. Number of edges in $N(\gamma)$ of the dataset of the bacterial genome of *X. fastidiosa*.

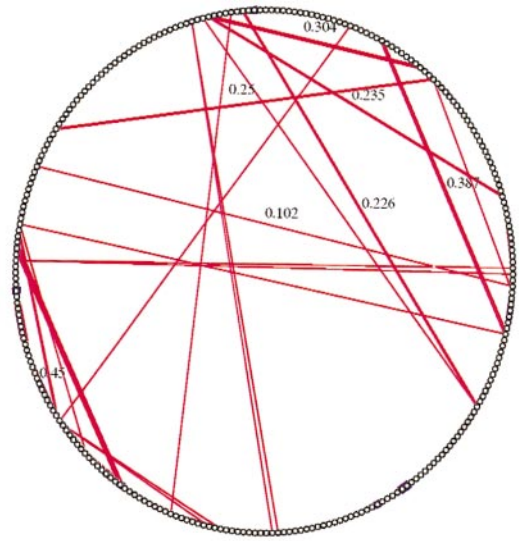
hood increases, slowly finding all true edges although gradually including more and more false edges. The curve shows a steep ascent until a γ of around 0.95. At that point $N(\gamma)$ contains about 360 edges including 98.5% of the 201 true edges. Remarkably, the size of $N(\gamma)$ increases by leaps and bounds between the confidence levels $\gamma = 0.95$ and $\gamma = 1.0$. Here the gain of 1.4% (corresponding to 2.8 edges) of the edges of the true solution that are missing in the original solution has to be paid for by 388.5 additional edges in the confidence neighborhood. It seems unlikely that, in practice, these edges could be found, and therefore, we recommend using the confidence neighborhood only for a confidence level $\gamma \leq 0.95$.

Application to X. fastidiosa Data

We applied our algorithm to a dataset from the bacterial genome of *X. fastidiosa* (1053 clones, 270 probes), which was produced by Frohme *et al.* (unpublished). We used a resampling rate of $B = 1000$. In Fig. 5, we plot the size of $N(\gamma)$ against the confidence level γ. The shape of this curve is similar to our simulations, except that the size of N is on average 1.4 times larger. Figure 6 shows GB restricted to the edge set $N(0.95)$. The probes are arranged in a circle corresponding to the original solution. The bootstrap values of the edges are translated into the edge width. Edges with a bootstrap value less than 0.1 are hidden. The remaining chords of the circle correspond to potential candidates for serious errors in the map, which can influence the "global structure" of the probe order. We also show an enlargement (Fig. 7) of Fig. 6 at a "weak" point in the original solution. The edge $e = (53, 75)$ has a small bootstrap value $b(e)$ of 0.185, and probe 53 shows connections to remote probes. To increase the confidence in this part of the solution, one would recommend further experiments.

DISCUSSION

Most physical mapping algorithms compute a single probe order as the solution of the physical mapping problem. In contrast, we focus on the determination of parts of the solution that are well defined by the data, as opposed to regions that are ambiguous. This allows us to focus further attention on ill-defined regions and to perform there additional experiments. Our approach is based on the bootstrap method, which has become



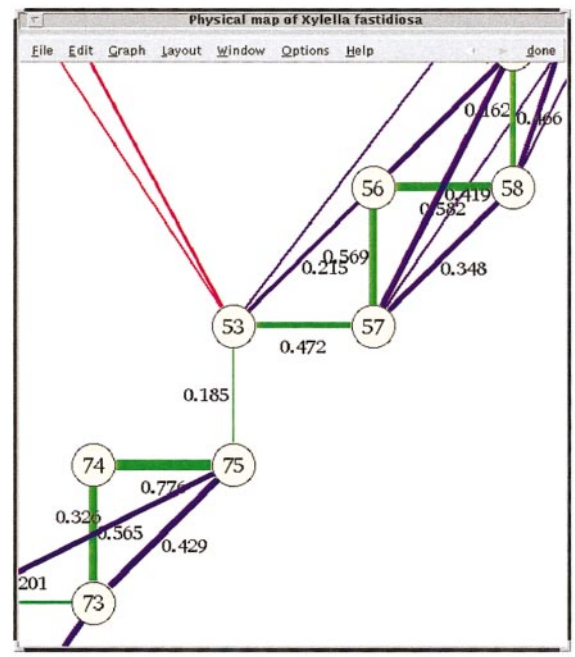


FIG. 6. A visualization of GB restricted to the edge set $N(0.95)$ of a dataset of the bacterial genome of *X. fastidiosa*. The nodes of GB are arranged corresponding to π. The edge width corresponds to the bootstrap value (for orientation a few edges are annotated with their bootstrap values), edges with bootstrap value less than 0.1 are not shown.

FIG. 7. Enlargement of Fig. 6. The edges that belong to $E(\pi)$ are colored green, edges that connect nodes with rank difference in π smaller than 10 are colored blue, the remaining edges are colored red.

one of the major tools for producing empirical confidence intervals of estimated parameters. We use this tool to measure the reliability of putative probe configurations in a physical map. Bootstrapping was also used by Wang *et al.* (1994), although with the role of clones and probes interchanged. Our method has the advantage that, due to the higher redundancy of the clones, the order of the probes is much more clearly defined by their hybridization pattern. Furthermore, we can rely on many more independent data points leading to more reliable estimations.

In general, bootstrap values do not necessarily correlate with accuracy—see Hillis and Bull (1993) for an empirical study of this question in the context of phylogeny. In physical mapping, there are three main reasons that could cause a decrease in the bootstrap value of a true edge: First there could be ambiguity caused by nonuniqueness of the solution. This ambiguity is present even under idealized, error-free conditions and could be encoded in the PQ-tree structure of the solution. Alternative reasons for a low bootstrap value could be low clone coverage or noisy data.

In our experimental setting, a simulation study (see Correlation of Bootstrap Values and Error Rate) suggests that the bootstrap values are good estimators of the probability that two probes occur consecutively in the true probe order. This good correspondence may be in part due to the high clone coverage we assume.

Based on bootstrap values, we construct a neighborhood N of alternative edges to the original solution. Our definition of N mimicks a confidence interval based on the percentile method described by Efron and Tibshirani (1993). This has the advantage that only a small number of highly likely alternative configurations is reported. In the parts of the original solution in which the probe order is well supported by the data, the neighborhood N contains no additional edges, while in ill-defined regions it contains "reasonable" alternatives that occurred in a high percentage of the bootstrap replicates. This highlights the regions of low quality and simultaneously offers the possibility of performing additional experiments, reducing the ambiguity.

The edges of N not part of the original solution can be partitioned into a set of edges that connect probes with small rank distance in the original solution and a set of edges with large rank distance (by a given threshold). While the first set corresponds to local ambiguity in the map, which is in practice of only minor interest, the latter edges may be signals of serious errors (see also Mayraz and Shamir, 1999, for a similar definition). This leads to a further decrease in the size of the candidate set for reasonable probe orders.

Although our method performs well in our simulation study as well as for the *X. fastidiosa* genome, it remains to be seen how well it generalizes for other physical mapping techniques and more problematic targets. In the complex work flow of physical mapping procedures there are many possible sources for correlated errors like contaminations, mix-up of clones, systematic measurement errors, chimeric clones, and complex repeats in eukaryotic DNA (Greenberg and Istrail, 1995; Hanke *et al.*, 1998). All these errors influence the performance of physical mapping algorithms, but in their entirety they are hard to model by an objective function or to test by simulations. This is the reason physical mapping algorithms typically include data preprocessing steps to handle such systematic errors. It was our goal to present here a simple and practical tool that can easily be combined with any of these methods and that we hope complements their performance by drawing attention to the remaining ambiguity in the physical map.

In the future, we intend to generate alternative (local) probe orders that use the edges in the confidence neighborhood as a candidate set. Such orders could be used for evaluation by more complicated objective functions similar to the bootstrap "bumping" strategy (Tibshirani and Knight, 1997), for detection and elimination of inconsistent hybridization signals, and for automatic selection of additional probes from ill-defined regions or contig ends. We also plan to adapt our method to STS-content data.

ACKNOWLEDGMENTS

We thank Jens Stoye, Richard Desper, Jens Hanke, and Marcus Frohme for many helpful discussions. Marcus Frohme provided us with the data set of *X. fastidiosa*.

REFERENCES

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

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.

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.

Daly, M., Reeve, M., Kaufman, A., Orlin, J., and Lander, E. (1994). CONTIGMAKER: Software for physical map contig assembly. Cold Spring Harbor meeting on genome mapping and sequencing, Cold Spring Harbor, NY, p. 210.

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

Efron, B., and Tibshirani, R. (1993). "An Introduction to the Bootstrap," Chapman & Hall, New York.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783–791.

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* chromosomes III and IV. *Electrophoresis* **19:** 482–485.

Hillis, D., and Bull, J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42:** 182–192.

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, L., Jr., 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.

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, Boca Raton, FL.

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

Melhorn, K., and Näher, S. (1995). LEDA: A platform for combinatorial and geometric computing. *Commun. ACM* **38:** 96–102.

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., 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 Univ. 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. *Nucleic Acids Res.* **23:** 3842–3849.

Setubal, J., and Meidanis, J. (1997). "Introduction to Computational Molecular Biology," PWS, Boston.

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

Tibshirani, R., and Knight, K. (1997). Model search and inference by bootstrap "bumping." Technical report from the Department of Statistics, University of Toronto. Presented at the Joint Statistical Meetings, Chicago, August 1996.

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.

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.