LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

LA-UR--90-2373

DE90 014916

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE COMPUTATIONAL METHODS FOR PHYSICAL MAPPING OF CHROMOSOMES

AUTHOR(S) DAVID C. TORNEY
CLIVE C. WHITTAKER
STEVEN W. WHITE
KAREN R. SCHENK

SUBMITTED TO For publication in Proceedings of Conference on Electrophoresis, Supercomputing and the Human Genome, Tallahassee FL, April 10-13, 1990. Tallahassee: Florida State University.

DISCLAIMER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

COMPUTATIONAL METHODS FOR PHYSICAL MAPPING OF CHROMOSOMES*

by

David C. Torney

Karen R. Schenk

Theoretical Biology and Biophysics (T-10)
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545 U.S.A.

Clive C. Whittaker

IBM New Mexico and Theoretical Biology and Biophysics (T-10)
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545 U.S.A.

Steven W. White

IBM New York
Dept. 4KJ, MS 601
Kingston, NY 12401

TEX: PR.cm

* Work performed under the auspices of the U. S. Department of Energy

Computational Methods for Physical Mapping of Chromosomes

First International Conference on Electrophoesis Supercomputing and the Human Genome. Tallahassee, FL, April 10-13, 1990

David C. Torney, Clive C. Whittaker, Steven W. White, and Karen R. Schenk

Los Alamos National Laboratory, T-10, MS K710, Los Alamos, NM 87545 (DCT, KRS), IBM New Mexico, Los Alamos National Laboratory, T-10, MS K710, Los Alamos, NM 87545 (CCW), and IBM, Dept. 4KJ, MS 601, Kingston, NY, 12401 (SWW)

A standard technique for mapping a chromosome is to randomly select pieces (with replacement), to use restriction enzymes to cut these pieces into (sequence specific) fragments, and then to use the fragments for estimating the probability of overlap of these pieces. (Overlapping pieces are likely to "share" fragments).

Typically, the order of the fragments within a piece is not determined, and the observed fragment data from each pair of pieces must be permuted $N1! \times N2!$ ways to evaluate the probability of overlap. $N1$ and $N2$ being the observed number of fragments in the two selected pieces. We will describe computational approaches used to substantially reduce the computational complexity of the calculation of overlap probability from fragment data. Presently, about 10^{-4} CPU seconds on one processor of an IBM 3080 is required for calculation of overlap probability from the fragment data of two randomly selected pieces, with an average of ten fragments per piece. A parallel version has been written using IBM clustered FORTRAN. Parallel measurements for 1, 6, and 12 processors will be presented.

This approach has proven promising in the mapping of chromosome 16 at Los Alamos National Laboratory. We will also describe other computational challenges presented by physical mapping.

Introduction

One begins physical mapping by fingerprinting with a library of cloned pieces from the *target*, the region to be mapped. Each cloned piece is then "fingerprinted" cut into fragments (using restriction enzymes) and the repetitive sequences present

in each fragment can be determined[1]. In some fingerprinting strategies the terminal sequence of a restriction fragment is determined[2]. This paper addresses how fingerprint data can be used to find overlapping cloned pieces. Clearly, the more apparently identical restriction fragments are shared by two cloned pieces, the greater the chance of overlap. More shared fragments will be required if, as is typical, the order of the fragments within the cloned piece is not determined (than if the order were known).

In this article we describe computer algorithms developed to determine the probability of overlap of two cloned pieces given fingerprint data, when the order of the restriction fragments is not determined. Before describing these algorithms, we will summarize the formulas that are evaluated.

Clone Overlap Probabilities

These statistical considerations are described more generally elsewhere[3].

Overlap probabilities would be optimally determined by the likelihood of the fingerprint data of two clones and overlap, and the likelihood of the fingerprint data and nonoverlap - using a reasonable statistical model and Bayes' formula.

$$P(\text{overlap}|\underline{S}) = p(\underline{S} \text{ and overlap})/p(\underline{S}) \ ;$$

$$p(\underline{S}) = p(\underline{S} \text{ and overlap}) + p(\underline{S} \text{ and nonoverlap}) \ . \qquad (1)$$

Equation 1 is true regardless of what the variable $\underline{S}$ represents, but the optimal discrimination of overlap follows if one identifies the fingerprint data with $\underline{S}$.

If there is more than one restriction digestion in the fingerprint data, and if the separate digest fingerprints are not independent, it would be possible to derive Eq. 1 with $\underline{S}$ equal the fingerprint data only if that data affords a restriction map or a small set of possible restriction maps. The Los Alamos fingerprint protocol[1] currently uses three complete digestions: two digestions with one enzyme and a double digestion with the same two enzymes. Thus these three digests are manifestly not independent. Furthermore, even if the digests were all generated with different enzymes, these fingerprints would still be dependent because the repetitive sequences present in the clone manifest themselves on fragments in all digests. In any case, noise in the fingerprint data makes it unlikely that one can reliably construct restriction maps separately for each clone.

Since it is apparently not practical to identify the Los Alamos fingerprint data with $\underline{S}$ in Eq. 1, the following approach was developed. We let

$$\underline{S} = \{S_E, S_H, S_{EH}\} \ , \qquad (2)$$

be an overlap statistic with three components, the latter being the likelihood ratio of the fingerprint data of one digest in two clones and overlap to that data and

nonoverlap, using an appropriate statistical model; S_E is derived from the EcoR1 digests; S_H is derived from the Hind3 digests; and S_{EH} is derived from the double digests with the same enzymes. We will proceed to write formulae for the components S of $\underline{S}$.

The main assumptions of the statistical model are that the restriction sites and hybridizing repetitive sequences are randomly (uniformly) and independently placed - reasonable assumptions based on the Los Alamos fingerprint data.

An indication of this is seen in Fig. 1, showing the fragment size distribution in the three digests of 2,200 clones. Except for fragments smaller than 1 kilobase in size, the fragment size distributions are exponential, consistent with random (uniform) placement of restriction sites, with no obvious contributions from any fingerprint fragments (due to repetitive DNA sequences) repeated throughout the target. Another important assumption concerns the noise in the data. We assume the fragment sizes are measured with a Normally distributed component of noise, proportional to the fragment size.

Figure 2 shows this assumption is consistent with the data, based on approximately 30,000 pairs of measurements of fragments likely to be the same fragment in overlapping clones. A least-squares fit gives the standard deviation of the noise equal 0.005 multiplied by the fragment size. Since noise in the hybridization data is at a low level, it is ignored to a first approximation; but this can readily be included[3] in S.

To compute S, one begins with a matrix $\underline{C}$ with matrix elements:

$$c_{ij} \equiv \frac{H_{GT} \cdot H_{RS} \cdot \ell_r \cdot \exp\{(\ell_{1i} + \ell_{2j})/2\ell_r\} \cdot \exp -\{(\ell_{1i} - \ell_{2j})^2/2\epsilon^2(\ell_{1i}^2 + \ell_{2j}^2)\}}{\epsilon\sqrt{2\pi(\ell_{1i}^2 + \ell_{2j}^2)}} \quad . \tag{4}$$

In Eq. 4, ℓ_{1i} is the length of restriction fragment from the first clone, ℓ_{2j} is the length of restriction fragment j from the second clone, ℓ_r is the average length between restriction sites, and ϵ times the length of a fragment is the standard deviation of length measurement reproducibility; ϵ equals 0.5%. Also, H_{GT} and H_{RS} are factors reflecting results of hybridization to GT repetitive sequence and Repetitive Sequence (Cot1) probes. These H are a function of λ, the ratio of the average length of compared fragments to the average distance between occurrences of the corresponding hybridization site. If both fragments hybridize, H is $\exp(\lambda)$; if neither fragment hybridizes, H is $[1 - \exp(\lambda)]^{-1}$; otherwise H is 0. Naturally, most c_{ij} are negligible.

S is derived from $\underline{C}$ as follows:

$$S = \sum_{k=1}^{\min(N_1, N_2)} \sigma_k \quad . \tag{5}$$

where

$$\sigma_k = \frac{(N_1 - k)!(N_2 - k)!}{N_1!N_2!} \sum_{i_1,i_2,\ldots,i_k=1}^{N_1}{}' \sum_{j_1,j_2,\ldots,j_k=1}^{N_2}{}' \prod_{\ell=1}^{k} c_{a_\ell j_\ell} \; ; \; N_1, N_2 \geq k \quad .$$

N_1 and N_2 are the number of fragments of the two cloned pieces. The primes on the summation signs in Eq. 6 indicate that no two summands are equal.

The computational challenge is to effectively evaluate Eq. 5, and some preliminary algorithms are described in the next section.

Computer Algorithms

To compute the matrix $\underline{C}$, one begins by sorting the fingerprint data according to ascending fragment size. This facilitates computing only those C_{ij} that are above a threshold. (Since C_{ij} is dominated by the Gaussian, only fragments with sizes within a "window" need be compared with a given fragment; and uncomputed matrix elements are taken equal zero).

At the next stage in the calculation, Eqs. 4 and 5, the sum of all possible products of n matrix elements (with no more than one element from any row or column of the matrix in any product) must be computed. The matrix is now reduced by extracting all nonzero elements for all other elements in its column and row and then deleting the column and row. The matrix is further reduced by extracting the sum of all nonzero elements in a column (row) with zero for all the other elements in the nonzero elements' rows (columns) and then deleting the column (row) and rows (columns). The reason for this reduction is that these extracted elements and sums of elements can be used in products independently of one another. To calculate the sum of products of n elements, in Eq. 5, one can take n' from the extracted elements and sums of elements, and $n - n'$ from zero to n. Using recursion, it is possible to compute the sum of products of extracted elements taken one at a time through n at a time in a number of operations proportional to n^2. For the residual matrix, with elements that cannot be chosen independently, algebraic manipulations were performed to greatly reduce the complexity. For example, consider

$$T_2 = \sum_{i,j,i',j'=1}^{M,N}{}' C_{ij}C_{i'j'} \quad , \tag{6}$$

where the prime on the summation indicates that i cannot equal i' and j cannot equal j', M being the upper limit for i and N being the upper limit for j. This can be rewritten:

$$T_2 = \sum_{i,j,i',j'}^{M,N} C_{ij}C_{i'j'} - \sum_{i,j,j'=1}^{M,N} C_{ij}C_{ij'} - \sum_{i,i',j}^{M,N} C_{ij}C_{i'j} + \sum_{i,j}^{M,N} C_{ij}^2 \quad . \tag{7}$$

Each of these four terms can be evaluated with the number of operations proportional to $M \times N$. The complexities of T_3 and T_4, evaluated in analogy with Eqs. 6 and 7, are $M \times N$ and $M^2 \times N$ (or $M \times N^2$), respectively. In preliminary versions of our programs, we do not take sums of products of more than four elements from the reduced matrix. Although this truncation has no effect on the accuracy of overlap detection for fingerprint data generated at Los Alamos, we are exploring techniques for efficient evaluation of the reduced matrix so that the algorithm would be useful for fingerprints with many similar restriction fragments in a typical clone. In this situation, one must address how well the experimental technique reveals the multiplicity of near-identical fragments.

Simulations

The probabilities appearing on the right of Eq. 1 are evaluated by Monte Carlo simulation of nonoverlapping or overlapping pairs of clones in FORTRAN programs FALSE and TRUE run on an IBM 3090 computer. The parameters of the simulation were chosen so that selected features of simulated clones were very similar to those observed in the data. Normal "noise" with standard deviation $\epsilon \times 1$ is added to a restriction fragment of length 1, modeling the reproducibility of apparent length measurement in our experiments. This noise can be decomposed into noise that is correlated for all fragments in a clone fingerprint, and noise that is uncorrelated with the latter dominant. To model GT nucleation, GT hybridization sites were randomly placed with the given average spacing and clones randomly selected, not containing at least one GT site, are rejected. To model the nondetection of small GT negative fragments less than 1.2kb in length, these were discarded if less than 500 bases; otherwise they were kept with a probability equal to: (length-500)/(1200-500).

The integer part of the logarithms of the three statistics is used to construct (three-dimensional) histograms of the outcomes of the simulations of nonoverlapping and overlapping pairs. Cubic interpolation from the 64 nearest "bin" coordinates is used to evaluate Eq. 1 for arbitrary $\underline{S}$. Typically, 5×10^7 simulated pairs of overlapping clones and 10^9 simulated pairs of nonoverlapping clones are more than adequate for subsequent data analysis. It takes approximately 3×10^{-4} cpu seconds on one processor of the IBM 3090E to evaluate Eq. 1 for a randomly selected pair of clone fingerprints. The formulas discussed in this manuscript and the computer program used to evaluate them can be generalized to encompass fingerprint strategies based on fragments whose order is not known. Similar formulae apply if restriction maps are known for the clones, but the computational complexity of overlap detection would be substantially smaller.

Parallelization

Results were announced at the conference on parallelizing a version of the FORTRAN program FALSE using clustered FORTRAN hardware and software installed on a pair of ES/3090 600Js with 12 Vector Facilities. These results are summarized in Table 1; more detail is presented elsewhere[4]. The substantial parallelization achieved with this program could easily be achieved in current and planned versions of FALSE and in programs used to analyze data.

Table 1

Number of Clones	Processors	First-to-Last User Instruction Speed-up	Complete Application Speed-up
2000	1	1.00	1.00
	6	5.71	5.03
	12	10.93	6.08
4000	1	1.00	1.00
	6	5.77	5.60
	12	11.44	9.45
8000	1	1.00	1.00
	6	5.81	5.77
	12	11.78	11.13

Results and New Directions

Some results from the Los Alamos clone mapping protocol and the analysis described above are illustrated in Figs. 3 and 4. Figure 3 depicts a histogram of the number of clone pairs determined to have overlap probabilities between 0.1 and 1.0 when approximately 2,200 (mostly) GT nucleated cosmid clones from chromosome 16 were fingerprinted. The expected number of (the?) clone pairs with overlap probability > 0.01 is 2,935; whereas, 2,750 is predicted from the GT nucleated prior probability. This slight excess can be explained on the basis of centromeric repeat fingerprint motifs present in about 55 nonoverlapping clones.

Figure 4 contrasts the efficacy of overlap detection for some variations in the fingerprint protocol. Fingerprint data was simulated using our statistical model

with parameters from the Los Alamos experiments. The plot shows the proportion of overlaps detected (essentially the detection probability) against the proportion of the clones that is shared. Here, we define overlap to be detected when the posterior overlap probability exceeds 0.5. From the plot, we see that half the overlaps are detected when the shared proportion is 0.4 using the most informative fingerprint, with three digests and three hybridization probes. An overlap fraction of 0.55 is required for 50% detection for the three digests and no hybridization fingerprint.

Clone overlap detection is necessary but not sufficient for completion of physical maps. Statistical methods are under development to determine the robustness of *contigs*, overlapping sets of clones, and to reduce these into maximally likely spanning sets that would serve as starting materials for sequencing.

Acknowledgements

We thank George Bell and Thomas Marr for encouraging comments at the early stages of this project. David Balding has collaborated on the development of the statistical models described above. Doyce Nix of IBM helped in converting these routines for the 3090 600 computer, and IBM provided a 3090 600 for code development and studies of parallelization as part of the IBM/Los Alamos Joint Study.

This work was done in the Center for Human Genome Studies at Los Alamos National Laboratory and was supported by United States Department of Energy grant # B04861/F118 (?? Not sure I read this right).

References

1. R. L. Stallings, D. C. Torney, C. E. Hildebrant, J. L. Longmire, L. L. Deaven, J. H. Jett, N. A. Doggett, and R. K. Moyzis, *PNAS USA* (1990), in press.

2. S. Brenner and K. J. Livak, *PNAS USA* **86** (1989) 8902-8906.

3. D. J. Balding and D. C. Torney, *Bull. Math. Biol.* (1990), submitted.

4. S. W. White, D. C. Torney, and C. C. Whittaker, *Supercomputing '90* (1990), submitted.

Figure Captions

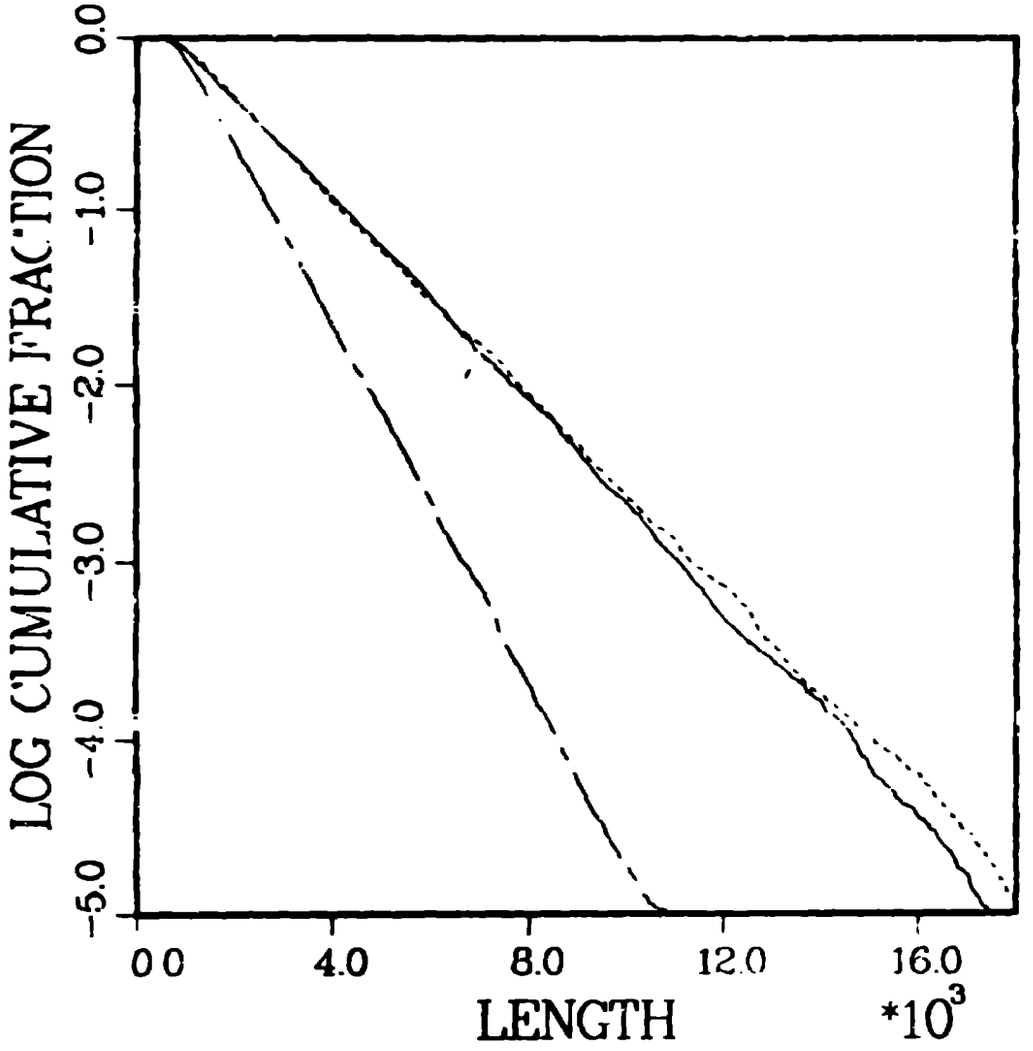
Figure 1. Fragment length distribution for double digest and two single digests (Los Alamos clone mapping protocol). Length in base pairs.
EcoR1 single digest (solid)
HindIII single digest (dash)
EcoR1 and HindIII double digest (doubly-dashed)

Figure 2. Fragment discrepancy histogram. Discrepancy is defined to be $(x - y)/(x^2 + y^2)$, x and y being two length measurements. Fragment pairs likely to be the same fragment were identified in clone pairs with inserts overlapping with probability > 0.9, using approximately 2,200 fingerprinted GT nucleated clones. The standard deviation of the discrepancies is found by doing a least-squares best fit of a Gaussian curve plus a baseline to the histogram; the standard deviation is found to be .005.

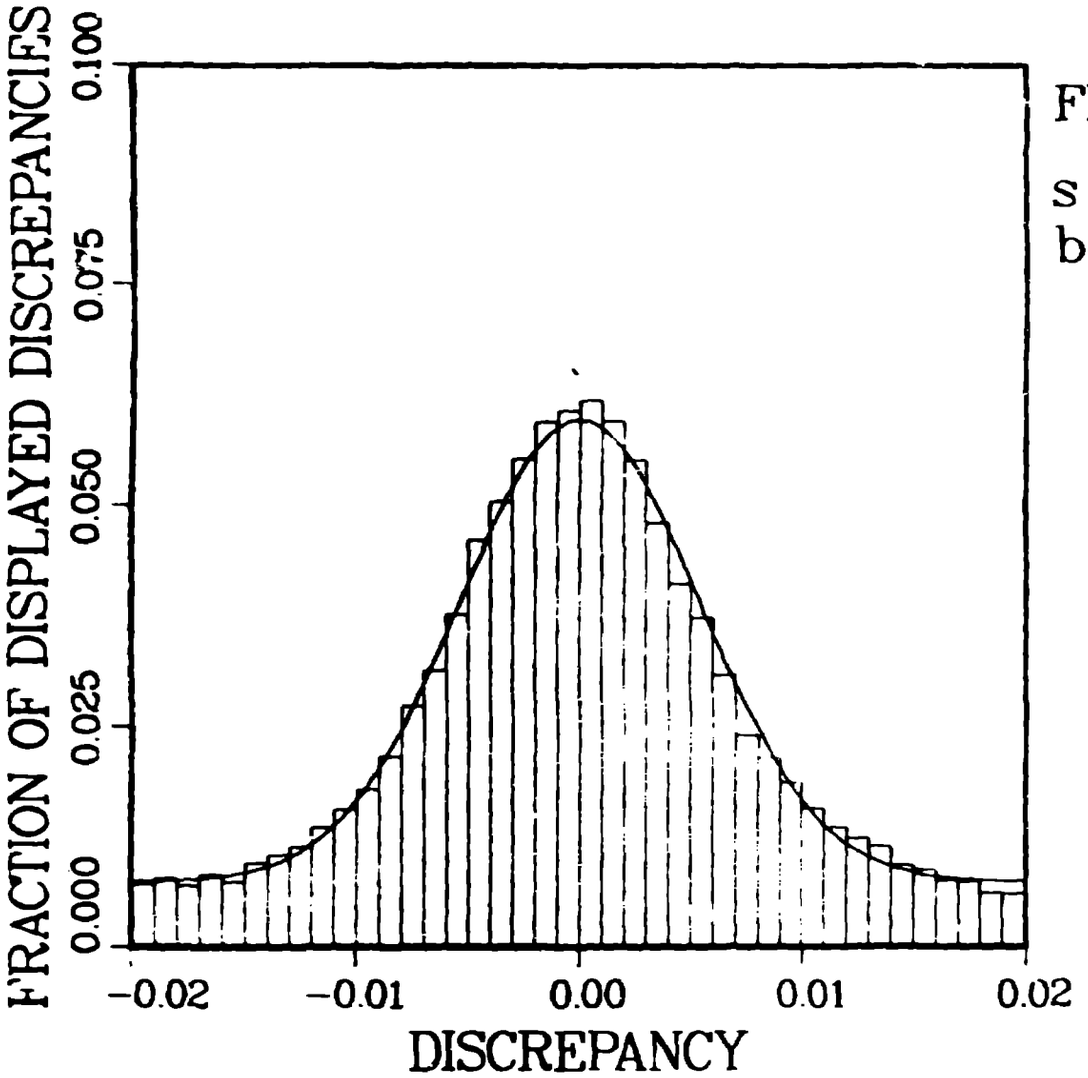
Figure 3. Posterior overlap probabilities ≤ 0.1 after 2,200 clones had been fingerprinted. Overlap probabilities calculated according to method described in Section 4 with data from three restriction digests and two hybridization probes. Most clones were selected on the basis of GT nucleation.

Figure 4. Comparison of the information available from different fingerprints: two single digests only with two or three hybridization probes and three digests (two single; one double) with zero, two and three hybridization probes. Noise is added to the simulated data to make it resemble real data. (Fragments are not detected and Normally distributed length measurement errors are added with a standard deviation of .008.)

FRAGMENT SIZE DISTRIBUTION



FRAGMENT DISCREPANCY HISTOGRAM

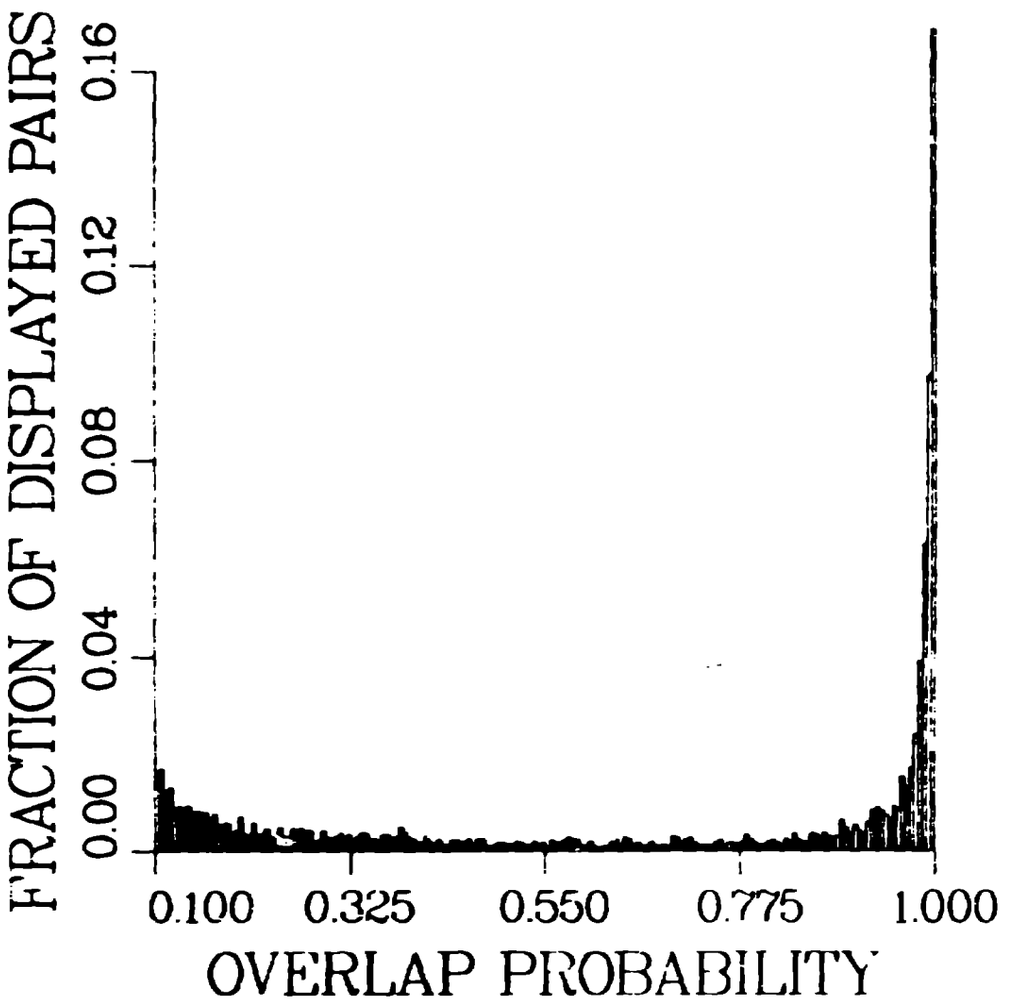


FIT PARAMETERS:

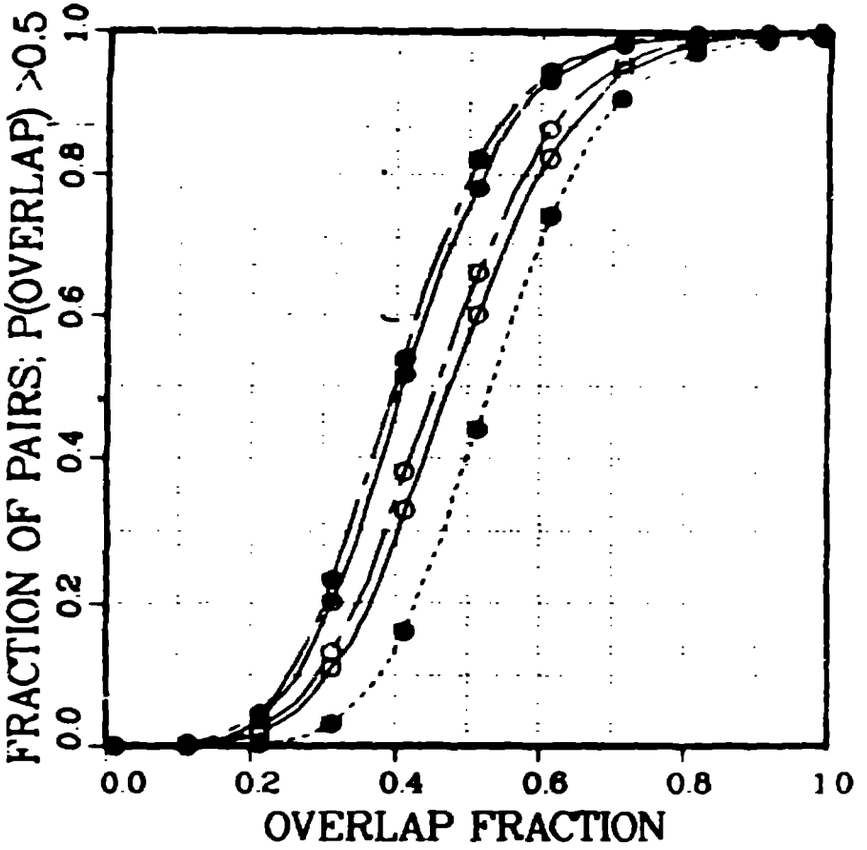
s = 0.005367

b = 0.293727

CLONE PAIR PROBABILITY HISTOGRAM



PAIR DETECTION: GT NUCLEATION



FINGERPRINT LEGEND

TWO HYBRIDIZATIONS:
GT(1/40KB) AND COT1(1/5KB)

THREE HYBRIDIZATIONS:
GT AND COT1 AND L1(1/60KB)

NO HYBRIDIZATIONS

TWO SINGLE AND
ONE DOUBLE DIGEST

TWO SINGLE DIGESTS