Data and Algorithms for Genomic Physical Mapping

by

Alan P. Kaufman

Submitted to the Operations Research Center
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1994

Author .....
Operations Research Center
July, 1994

Certified by .....
James B. Orlin
Professor of Operations Research
Thesis Supervisor

Accepted by .....
Richard C. Larson
Professor of Electrical Engineering
Codirector, Operations Research Center

Data and Algorithms for Genomic Physical Mapping

by

Alan P. Kaufman

Submitted to the Operations Research Center
on July, 1994, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

This thesis presents the first independent assessment of two physical mapping projects: the CEPH-Genethon fingerprint mapping effort and the CEPH-Genethon ALU-PCR mapping effort.

The fingerprint data are found to contain numerous errors. Three novel statistics are developed to use these data to determine overlapping pairs of CEPH-Genethon YACs. The best of these statistics is of comparable power to the more sophisticated CEPH-Genethon LOS measures. One novel statistic has proved useful in resolving ambiguous YAC-STS addresses, with concomitant savings in laboratory time and resources.

The ALU-PCR data and their accompanying map construction strategy generate a map with numerous errors. In particular, this strategy treats one-third of the ALU-PCR probes as "wild-card" probes, valid on any chromosome. The CEPH-Genethon strategy applies a single-copy probe mapping algorithm to multiple-copy probes. The resulting map is riddled with spurious connections. An improved map construction strategy is developed using insights from graph theory.

Thesis Supervisor: James B. Orlin
Title: Professor of Operations Research

Acknowledgments

I wish to thank Professor Jim Orlin for his support, patience, and advice.

I wish to thank my family for their support and encouragement.

Finally, I wish to thank Sara Elisabeth. With her, every day is wonderful, and life is a great party-on-wheels.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Physical maps of the human genome order landmarks and DNA fragments along the human chromosomes. Such maps are invaluable tools in the battle against human genetic diseases. Different strategies exist for developing such maps. This thesis examines two strategies: overlap detection via restriction enzyme fingerprints and overlap detection via hybridization probes.

Chapter two reviews the literature on fingerprint mapping. Chapter three addresses the mathematics of fingerprint-based tests for determining pairwise clone overlap. Chapter four applies these tests to real and simulated data and investigates their implications for contig construction.

The last chapter examines probe mapping methods. Chapter five reviews the CEPH-Genethon ALU-PCR mapping project. The chapter highlights problems in the CEPH-Genethon map and proposes some partial remedies.

This thesis focuses on overlap detection, mapping algorithms, and data assessment. Except where relevant to the mathematics, detailed explanations of underlying biological mechanisms are generally avoided. Some biological terms are defined briefly in Appendix A. Daggers† accompany their first occurrence in the text. The reader may consult the excellent overview of the Human Genome Project [16] or a recent, comprehensive masters thesis [19] for more details.

Chapter 2

Fingerprint Mapping Literature Review

This chapter reviews recent mapping projects, strategies, analysis methods, and algorithms involving restriction enzyme† fingerprint patterns.

Before 1986, restriction fragment mapping projects had covered small regions, 50-100 kilobases† in size. The methods of these projects were not suitable for larger regions.

In 1986, two mapping projects simultaneously attempted a radical new strategy: fingerprinting libraries† of randomly created clones†. This method allowed the mapping of much larger regions, and proved successful for the 15 megabase† genome† of the yeast *Saccharomyces* and the 80 mb genome of the nematode worm *Caenorhabditis elegans.* In 1987, a similar approach mapped the 4.7 mb genome of the bacterium *Escherichia Coli.* The first mathematical analysis of fingerprint mapping appeared in 1988. In 1990, the strategy was applied to Human Chromosome† 16, 90 mb in length. The most ambitious use of the method occurred in 1992: a massive fingerprinting effort on a random clone library in an attempt to map the entire Human Genome, 3300 mb in length.

Differing in their digestion methods, analysis techniques, and size, these projects enjoyed varying levels of success. Following a section on Bayesian assessment of overlap data, subsequent sections of this chapter examine the salient features contributing

to the success or failure of these projects.

2.1 Bayes Law and Overlap Detection

One may use Bayes Law to write the probability that two clones overlap conditional on observing some data, D, and the prior probability of overlap, p_{OL}. To be precise, overlap is a continuous characteristic. Defining a minimal overlap threshold converts this continuous quantity into a binary result. The following basic results consider overlap as a binary characteristic. Continuous overlap models are introduced in Section 2.6.2.

$$P(\text{ OVERLAP}|D) = \frac{P(D|\text{ OVERLAP})}{P(D)} p_{OL} \tag{2.1}$$

This may be written in terms of a likelihood function, $L(D)$.

$$L(D) = \frac{P(D|\text{ NO OVERLAP})}{P(D|\text{ OVERLAP})} \tag{2.2}$$

$$P(\text{ OVERLAP}|D) = \left(1 + \frac{1 - p_{OL}}{p_{OL}} L(D)\right)^{-1} \tag{2.3}$$

Equation 2.1 or Equation 2.3 represent the correct way to compute a posterior overlap probability from an observation and a prior.

2.2 The Yeast Genome

2.2.1 Experimental Method

In 1986, Olson et al. [43] created a library of 5000 λ clones, each containing an insert of yeast DNA. The average insert size was 15 kb, providing 5-fold coverage of the 15 mb yeast genome. Chimerism†, deletions†, or other cloning difficulties were not reported, reflecting the relative stability of the λ vector.

The 5000 clones were double-digested† with two restriction enzymes, EcoRI and HindIII. As no distinction was made between the two types of restriction sites, the

generic term "RH" was used to refer to the double digest cleavage sites. EcoRI and HindIII are both 6-cutters. Using the random-base DNA model†, which crudely models the four bases of DNA as equally probable, a given 6-cutter recognition site† occurs every 4^{-6} bases on average. The double digest cuts this distance by half. Thus, one expects each 15 kb λ clone to contain 7.3 RH sites, producing 8.3 RH fragments. The observed mean was 8.36 fragments.

Gel photographs were projected onto the surface of a digitizing tablet and manually traced. The raw images were converted to fragment sizes in basepairs† by polynomial interpolation against control bands of known size.

2.2.2 Pairwise Overlap Detection

Pairwise comparisons were made between pairs of fragment size lists. The two lists corresponded to the fragments of two clones, or of one clone and a partially built composite map. In the second case, the list corresponding to the partial map was designated the "reference list", and the list corresponding to the clone was designated the "comparison list." If both lists were single clones, the assignment of "reference" and "comparison" was arbitrary.

The yeast team did not adopt the Bayesian approach described in Section 2.1. Instead, apparent overlap between the pairs of lists was determined using a combination of statistical heuristics. First, each list was scanned independently for intra-list fragment identities. Adjacent bands falling within a thin "identity window" were merged into one. For a list corresponding to a clone, this operation removed doubly traced bands (data entry errors). For a list corresponding to a map, this operation produced a consensus fragment size from its multiple measurements. Next, similar bands between the two lists were paired if they fell within an "error window." Bands were not multiply paired. The width of this error window was expanded linearly with the size of the reference list fragment. This corresponds to a model of fragment measurement error with a standard deviation proportional to fragment length. The proportionality constant was not reported in [43].

The following notation is introduced to summarize the yeast project's clone overlap

rule. This notation is maintained throughout the thesis. Let $x_{a_1}, x_{a_2}, \ldots, x_{a_n}$ denote the sizes of the a_n fragments in the reference list and $x_{b_1}, x_{b_2}, \ldots, x_{b_n}$ denote the sizes of the b_n fragments in the comparison list. Let $m_{a_1}, m_{a_2}, \ldots, m_{a_n}$ indicate paired fragments in the reference list: $m_{a_i} = 1$ if the reference fragment i matches some fragment from the comparison list, and $m_{a_i} = 0$ otherwise. Indicator variables for the comparison list, $m_{b_1}, m_{b_2}, \ldots, m_{b_n}$, are defined analogously. Let $s = \sum m_{a_i} = \sum m_{b_i}$ denote the number of matched fragments. Let $d_1, d_2, \ldots, d_s$ denote the percent discrepancies of matched fragments, with mean $\bar{d} = \sum d_i/s$.

The yeast project overlap rule had four components:

Enough matches: $s > k_1$

Not too many mismatches: $\max(a_n - s, b_n - s) < k_2$

Mutual Overlap Statistic: $s^2/n_a n_b > k_3$

Adjusted Fit: $\sum(d_i - \bar{d})^2/s < k_4$.

Olson et al. considered an overlap significant when it satisfied these conditions with $k_1 = 4$ matches, $k_2 = \infty$ mismatches, $k_3 = 0.60$, and $k_4 = 1\%$.

2.2.3 Contig Assembly

Connected components in the clone-clone overlap graph provided preliminary, unordered contigs†. 85% of the clones fell into 680 contigs. The average contig size was 6.2 clones. Simulations indicated that an expected 10 false linkages would be generated by this overlap procedure, implying that an expected 10 of the 680 contigs linked unrelated sets of clones.

Topological constraints imposed by restriction fragment mapping were used to refine the preliminary contigs. Restriction maps† of the contigs were constructed with a greedy algorithm. An initial "seed clone" was selected. The best matching clone in the remainder of the contig was aligned against it, matching RH sites. Additional parsimonious clones were added to the alignment. Clones that did not fit the RH map

were removed from the contig. According to the yeast mapping team, this method removed all incorrect linkages in the preliminary contigs.

Restriction-mapped contigs were oriented and aligned using end clone fragments. The overlap conditions were relaxed so multiple weak relations between end clones could connect contigs.

2.2.4 Comment

The yeast project did not employ the correct Bayesian approach of Section 2.1 and did not justify their ad-hoc overlap test. However, this test was only used to generate preliminary contigs. Creating restriction maps ordered and verified each contig, improving map quality significantly.

According to the yeast mapping team. the final map covered 95% of the genome. Figure B-1 displays a portion of the yeast map.

2.3 The Worm Genome

2.3.1 Experimental Method

In 1986, Coulson et al. [17] amalgamated a heterogeneous library of cosmid and λ clones from various *Caenorhabditis elegans* research labs. The library contained about 8000 clones. The average insert size was 34 kb. This provided 3-fold coverage of the worm's 80 mb genome. (Additional cosmids, λs, and eventually YACS† were later added to the library, bring the total number of clones to over 17000 and the coverage to 18.[53]) It is interesting to note that this article, unlike the article announcing the yeast map [43], did not highlight these essential statistics. The relevant information is buried in the text and in figure captions. A mathematical analysis of fingerprint mapping had yet to be published[34]; thus, the analytical relationship between map quality and genome coverage, genome size, library size, and overlap detection sensitivity was unavailable.

The library was double digested with HindIII, a 6-cutter. The cut ends were

tagged and digested again with Sau3A1, a 4-cutter†. The lengths of tagged fragments were measured using electrophoresis†. Thus, most measured fragments corresponded to intervals of DNA flanked by a HindIII site on one side and a Sau3A1 site on the other. To be precise, a fragment could have been flanked by two HindIII sites. However, under the random-base DNA model, HindIII–HindIII intervals which lacked a Sau3A1 site were rare. Using standard results on competing poisson processes[21], the probability that a gap begun at a HindIII site terminated with a HindIII site was $\frac{4^{-6}}{4^{-6}+4^{-4}}$, or less than 0.06. The chance the gap terminated with a Sau3A1 site was over 0.94.

Thus, the number of fragments from a clone (roughly) equals twice the number HindIII sites on the clone. The expected number of fragments is

$$34 \text{ kb} \times \frac{1 \text{ HindIII site}}{4^6 \text{ bp}} \times \frac{2 \text{ frags}}{1 \text{ site}},$$

or 16.6 fragment per clone. Coulson et al. reported an average of 23 fragments per clone. This statistically significant discrepancy is consistent with larger inserts (47 kb) or a greater frequency of HindIII sites (1.5 times greater than the random-base rate of 4^{-6}).

After electrophoresis, gel bands were entered manually using a digitizing tablet or semi-manually by digitization with human confirmation. Bands were standardized against control bands of known size, but these measurements were left in mm and not converted to basepairs. Coulson et al. felt

> ...no useful information is served by [converting from gel measurements to bp estimates] because, in our strategy, the lengths of the fragments convey no information about the length of the clone. Furthermore, since the gels are denaturing there is no precise correlation between molecular size and position (although a given fragment will always run at the same position.)[53]

Neither chimerism nor deletions were reported in the clone library.

2.3.2 Pairwise Overlap Detection

The worm project did not adopt the Bayesian methodology of Section 2.1. Instead, they based their overlap calculation on $P(D|$ NO OVERLAP), which they termed $PROBCOINC$, for "probability of coincidence."[53]

As equations 2.2 and 2.3 indicate, the absolute size of $P(D|$ NO OVERLAP) is irrelevant. What matters is $L(D)$, the relative likelihood of $P(D|$ NO OVERLAP) to $P(D|$ OVERLAP). $L(D) < 1$ implies overlap is more likely than nonoverlap, and $L(D) > 1$ implies nonoverlap is more likely than overlap. Further, the significance of a "large" or "small" $L(D)$ value depends on the prior, p_{OL}.

Nonetheless, Coulson et al. used $PROBCOINC$ to determine pairwise clone overlaps. The notation of Section 2.2.2 is maintained. Without loss of generality, let $a_n \geq b_n$, so "reference list" refers to the clone with more bands and "comparison list" to the clone with fewer. Let L_{GEL} denote the length of the sequencing gel, in mm. Let the L_{TOL} denote the tolerance of the sequencing gel: band a_i can be matched to band b_j if $|x_{a_i} - x_{b_j}| < 2L_{TOL}$. This corresponds to fragment measurement error that does not vary across the gel.

Let p denote chance a band from the comparison list matched a band from the reference list.

$$p = P(m_{b_i} = 1) = 1 - \left(1 - \frac{2L_{TOL}}{L_{GEL}}\right)^{a_n} \tag{2.4}$$

Using the binomial probability mass function,

$$B(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2.5}$$

Coulson et al. defined $PROBCOINC$ to be probability of observing s or more matches:

$$PROBCOINC = \sum_{k=s}^{b_n} B(k, b_n, p). \tag{2.6}$$

Note that this model assumes that bands are independently and identically distributed uniformly across the gels. Also, Equation 2.4 allows the double matching of bands, even though the matching algorithm was not permitted to do so.

2.3.3 Contig Assembly

The worm mapping team used a semi-manual method for contig assembly. Considering pairs of clones with sufficiently small $PROBCOINC$ to be linked, connected components in the clone-clone overlap graph provided preliminary unordered contigs. Humans, assisted by a variety of subroutines which considered local structures in the clone-clone overlap graph, ordered and aligned these preliminary contigs. The subroutines assembled contigs in a greedy manner, starting with the most probable clone pair overlaps. Unlike the yeast mapping project, which used restriction mapping to exploit the higher discriminating power of many-to-many clone relations, the worm mapping project relied solely upon binary $PROBCOINC$ relations.

2.3.4 Comment

Like the yeast mapping project, the worm project did not employ the correct Bayesian approach of Section 2.1. Unlike the yeast project, the worm project did use a formal model of overlap. By considering only pairwise relations and ignoring band patterns, however, the worm project lost much of the resolution they might have obtained from restriction mapping.

According to Coulson et al., the final map covered 60% of the nematode's genome with 860 contigs, ranging in size from 35 to 350 kb. Figure B-2 displays a portion of the finished map.

2.4 The *Escherichia Coli* Genome

2.4.1 Experimental Method

In 1987, Kohara et al. [30] created a library of 3400 λ clones, each containing an insert of *E. Coli* DNA. The average insert size was 15 kb. This library provided 11-fold coverage of the bacterium's 4.7 mb circular, single-chromosome† genome. 1025 clones would later form the backbone of the map. This mapping set provided 3-fold genomic coverage.

The 3400 clones were partially digested with eight separate single digestions with eight different restriction enzymes: BamHI, HindIII, EcoRI, EcoRV, BgII, KpnI, PstI, and PvuII. These enzymes are all 6-cutters, and the random-base DNA model predicts an average of 3.6 sites for each in a 15 kb clone. A partial digest fragment can begin at any site, including the ends, and end at any site, including the ends. Rounding 3.6 up to 4, each partial digest produces about $\binom{4+2}{2} = 15$ different fragment lengths under the random-base model.[1]

Gel images were entered manually using a digitizing tablet. Chimerism was not mentioned, but deletions were indicated in some clones.

2.4.2 Pairwise Overlap Detection

Like the yeast and worm projects, the *E. Coli* team did not adopt the Bayesian methodology of Section 2.1. Like the yeast project, the *E. Coli* project did not calculate explicit overlap probabilities within its overlap test.

To determine overlap, the *E. Coli* team used only the relative order of the eight varieties of restriction sites. Fragment sizes were not involved in this calculation. Instead, the eight partial digests were run side-by-side on the same gel, and their relative order was determined in a manner analogous to the Sanger or the Maxam-Gilbert method for DNA sequencing[6].

As the *E. Coli* genome is 4.7 mb, the random-base model predicts an average of 1150 sites for each enzyme, or 9200 restriction sites in total. Thus, the order of these eight restriction sites on the *E. Coli* chromosome may be considered a 9200 symbol sequence written in an alphabet of eight symbols.

Assuming the eight varieties of restriction cleavage sites occur at random through-

[1]This is a back-of-the-envelope calculation, for $E(f(x)) \neq f(E(x))$. However, the correct value is quite close. As the chance of a restriction site beginning at any given base is low, the number of 6-cutter restriction sites on a 15 kb fragment under the random-base DNA assumption is well-modeled by a poisson random variable with mean 3.6. Numerical evaluation of this expectation,

$$E(\binom{x+2}{2}) = \sum_{i=0}^{\infty} \binom{i+2}{2} \frac{e^{-3.6} 3.6^i}{i!},$$

yields 14.7.

out the genome, the probability of observing a particular restriction site k-mer in a given location is 8^{-k}. For example, the chance that the order of the first six restriction sites on the genome is EcoRI, EcoRI, BglI, HindIII, BamHI, PvuII is $8^{-6} = 4 \times 10^{-6}$. The expected number of occurrences of this restriction site 6-mer across the genome is $8^{-6} \times 9200 = 0.035$. The number of occurrences of this (or any) restriction site 6-mer across the genome is well modeled by a poisson random variable of mean 0.035. Thus, the probabilities of 0, 1, and 2 occurrences of this 6-mer are 0.966, 0.034, and 0.0006, respectively. Conditional on one or more occurrences, the probability of exactly one occurrence is 0.982. The conditional probability of exactly two occurrences is 0.017. In short, any given restriction site 6-mer probably does not occur on the genome (probability 0.966). If a given 6-mer does occur, it probably occurs just once (probability 0.982). Following this logic, the *E. Coli* mapping project employed a simple test for overlapping clones: two clones overlap if they share six or more consecutive cleavage sites.

2.4.3 Contig Assembly

The *E. Coli* mapping project used the methods of multi-alignment shotgun sequence assembly ([6], [37], [44]) to order cleavage sites and clones. Once the correct order of sites had been determined, multiple observations of the same band were averaged to estimate inter-restriction site distances on the map.

2.4.4 Comment

The *E. Coli* mapping project, like the yeast and worm projects, did not employ the Bayesian approach of Section 2.1. The *E. Coli* multiple-alignment approach to contig construction imposed topological constraints and yielded good contigs. As this approach considers the relationships between multiple clones at once, this strategy more closely resembles the restriction mapping refinement stage in the yeast project than the simple pairwise relations used in the worm project.

According to Kohara et al., the final map covered 96% of the bacterium's genome

with 70 contigs, ranging in size from 20 to 180 kb. Figure B-3 displays a portion on the finished map.

2.5 The Human chromosome 16

2.5.1 Experimental Method

In 1990, Stallings et al. ([52] [54], [5], [58]) created a library of 26000 cosmid clones. Each contained an insert of human chromosome 16 DNA. The average insert size was 40 kb. The library was probed† with a $(GT)_n$ † probe, yielding 3145 $(GT)_n$ positive clones. Assuming 40kb clones[2], these provided 1.5-fold coverage of the 85 mb chromosome.

The 3145 clones were digested three times: two single digests with the six-cutters† EcoRI and HindIII, and one double EcoRI-HindIII digest. These digests were run out on gels and digitally scanned. Bands were detected by machine and converted to basepairs. The gels were also blotted onto membranes and probed with $(GT)_n$ and CotI repetitive sequence† probes. Thus, the following data were known for each band in each digestion of each clone: its length, its $(GT)_n$ hybridization† status (0 or 1), and its CotI hybridization status (0 or 1).

2.5.2 Pairwise Overlap Detection

The chromosome 16 project did employ the Bayesian approach of Section 2.1 to evaluate pairwise clone overlap probabilities. Instead of directly using the data, D, from a pair of clones, Stallings et al. substituted a statistic, $S = f(D)$. This statistic was applied to the three digests separately. Subscripts "E," "H," and "EH" refer to the EcoRI, HindIII, and EcoRI-HindIII digestions, respectively.

[2]Assuming $(GT)_n$ probes are rare and independent of clone-end digestion sites, clones containing probes will tend to be larger than average clones. Basic random incidence results[21] indicate the expected length of the 3145 clones was $E(L) + \frac{\sigma_L^2}{E(L)} = 40 + \frac{\sigma_L^2}{40}$, but the value of σ_L^2 was not reported in [52].

$$S = \{S_E, S_H, S_{EH}\} = \{f(D_E), f(D_H), f(D_{EH})\}.$$

A strategic choice of the statistic f would summarize the complexities of a pair of digestions with a single number, and do so with little loss of information regarding overlap or non-overlap of the pair. For this statistic, Stallings et al. selected a digest likelihood ratio.

$$S = f(D) = \frac{P(D|\text{ OVERLAP, simplifying model})}{P(D|\text{ NO OVERLAP, simplifying model})} \tag{2.7}$$

Equations 2.7 and 2.2 appear similar, but they are not. The digest likelihood ratio of Equation 2.7 is based upon a simplifying model of the digestions. The resulting S then plays the role of D in Equation 2.2. The final clone likelihood ratio, L from Equation 2.2, depends on the distribution of S for overlapping clones and the distribution of S for non-overlapping clones.

Stallings et al. adopted a complex statistic for f. This statistic involved a $n_a \times n_b$ matrix, C. An element of this matrix, denoted c_{ij}, represented the ratio of the probability x_{a_i} and x_{b_j} were two measurements of the same fragment to the probability that x_{a_i} and x_{b_j} were measurements of two different fragments. Thus C contained likelihood ratios for fragments. The "simplifying model" corresponds to the Lander-Waterman model, described in Section 2.8.

$$c_{ij} = \frac{H_{GT} \cdot H_{COT} \cdot \bar{l}_r \cdot e^{\frac{(x_{a_i}+x_{b_j})}{2\bar{l}_r} - \frac{(x_{a_i}+x_{b_j})^2}{2\epsilon^2({x_{a_i}}^2+{x_{b_j}}^2)}}}{\epsilon\sqrt{2\pi({x_{a_i}}^2+{x_{b_j}}^2)}} \tag{2.8}$$

This fragment likelihood ratio involved two parts. The first part, the H_{GT} and H_{COT} terms, reflected fragment likelihood ratio obtained by considering only the hybridization status of the two fragments. The remaining terms provided the fragment likelihood ratio obtained by considering the fragment lengths, where $\bar{l}_r$ denoted the average length between restriction sites and and ϵ denoted the standard deviation of

the length measurement reproducibility.

These fragment likelihood ratios were then combined, accounting for all ways to match the n_a fragments in one clone against the n_b fragments in the other:

$$S = \sum_{k=1}^{\min(N_1,N_2)} \frac{(N_1-k)!(N_2-k)!}{N_1!N_2!} \sum_{\substack{i_1,i_2,\ldots,i_k=1 \\ \text{no two indices equal}}}^{N_1} \sum_{\substack{j_1,j_2,\ldots,j_k=1 \\ \text{no two indices equal}}}^{N_1} \prod_{l=1}^{k} c_{i_l j_l}. \qquad (2.9)$$

The complexity of Equation 2.9 is equivalent to the computation of a permanent (a determinant with all subtractions replaced by additions[58]). This computation requires exponential time[55]. The chromosome 16 project computed this statistic for each digest for each pair of clones, a total $3 \times \binom{3145}{2} \approx 15 \times 10^6$ times. Parallel computation, efficient algorithms, and approximations [58] reduced the all-pairs running time to several hours.

Due to the complexity of this statistic, no algebraic probability density function for random vector $\{S_E, S_H, S_{EH}\}$ exists. Stallings et al. determined the density function of $\{S_E, S_H, S_{EH}\}$ given overlapping clones and given non-overlapping clones through massive simulations of model genomes.

2.5.3 Contig Assembly

Contigs construction occurred in the clone-clone overlap graph, where nodes represented clones and arcs represented overlap probabilities above a certain threshold of certainty. Chimerism was not addressed explicitly, though choosing a sufficiently high certainty threshold would remove overlaps between chimeric clone halves.

Without chimerism, Stallings et al. could order clones within contigs. They used interval graph techniques to coalesce contigs by lowering the overlap probability threshold. However, due to the limitations of contig assembly using pairwise overlap relations and "...the presence of repeated DNA sequences, [map construction] requires human intervention in various stages of constructing an ordered clone map from experimental data."[58]

2.5.4 Comment

After marveling at the complexity of Equations 2.8 and 2.9, one wonders if the fingerprint data quality warranted such intricate analysis. Lacking an algebraic probability density function, the behavior of $\{S_E, S_H, S_{EH}\}$ may only be understood through simulation. Sensitivity analysis is thus hindered. It is possible this statistic is driven essentially by the number of matching fragments in the two clones. Such simplifications would have been difficult to detect and confirm using simulation.

Stallings et al. report constructing 460 contigs from their 3145 cosmid clones, covering 54% of chromosome 16. The average contig size was 106 kb. Given the 1.5-fold coverage of the clone library, these are impressive accomplishments. The success of the project hinged upon probing the restriction fragments; this reduced the minimum detectable overlap considerably. (The effect of this reduction is addressed in Section 2.8.) Figure B-4 presents a portion of the finished map.

2.5.5 Other Applications of Chromosome 16 Fingerprints

Fickett and Cinkosky [22] used the Stallings et al. clone-clone overlap probabilities as data for a genetic algorithm† (GA) to determine good ordered contigs. They criticized the sequential greedy method used by the yeast and worm projects (Sections 2.2.3 and 2.3.3) the GA outperformed greedy methods on chromosome 16 data. They used three objective functions to evaluate clone permutations. Efficient horizons in this three-dimensional objective space imposed partial orders on proposed solutions. One objective involved the product of successive overlap probabilities; the second involved the estimated degree of overlap; the third involved the lengths of the clones and the chromosome. Fickett and Cinkosky's GA produced better contigs than those produced by the clone-clone overlap graph theoretic approach used initially. In one instance, the GA broke a contig generated by the earlier algorithm, and this correctness of this break was confirmed by FISH† mapping individual clones.

Soderlund et al. ([51],[50]) worked with the chromosome 16 restriction fragment directly, constructing restriction maps to order contigs. Again, better results were

reported than those obtained by the overlap graph approach. To build restriction maps, Soderlund et al. used a "noisy consecutive ones"[3] and heuristic search techniques. They coded their algorithms into an interactive graphical software package named GRAM for computer-assisted restriction mapping. Restriction maps were the focus of these efforts, not not for improving the chromosome 16 clone map.

2.6 The Human Genome

In 1992, Bellanne-Chantelot et al. ([9], [8], [32], [33]) attempted a daring experiment. CEPH-Genethon† attempted to map the entire 3300 mb human genome using random clone fingerprinting. Previously, the largest region upon which the method had been used was chromosome 16, at 85mb. Two innovations allowed the CEPH-Genethon team to scale up the method forty-fold: YACS offered significantly larger inserts than cosmids, and automated gel reading equipment speeded data entry.

2.6.1 Experimental Method

The CEPH-Genethon team created a library of 22000 YACS containing human DNA. The average insert size was 810 kb, providing 5-fold coverage of the genome. The YACS underwent three single 6-cutter digestions with the restriction enzymes EcoRI, PvuII, and PstI. After electrophoresis, the gels were blotted and hybridized for the Kpn repetitive sequence. Kpn-containing fragments were detected with chemiluminescence, scanned, digitized, and standardized against control bands of known size. (The library has since increased to 33000 YACS, of which 25000 have mean insert size of 1 mb. Another repetitive sequence probe has been added, THE. These additional data and their value are discussed in Chapter 3.)

[3] A binary matrix has the consecutive ones property if its rows may be permuted so that ones occur consecutively in all columns. The noisy consecutive ones problem seeks to minimize a function of the number of ones that must be changed to zeroes and the number of zeroes that must be changed to one to produce a matrix with the consecutive ones property.

2.6.2 Pairwise Overlap Detection

The CEPH-Genethon team did adopt the Bayesian framework of Section 2.1. Let l_a denote the length of the first clone and l_b the length of the second. Let θ denote the length of their common region of overlap. OVERLAP from Equation 2.1 corresponds to $\theta > 0$, and NO OVERLAP to $\theta = 0$. The prior on overlap, p_{OL}, is supplemented by a prior probability density function on θ, $\pi(\theta)$. As all degrees of overlap are a priori equally likely, the non-informative or flat prior was used for $\pi(\theta)$.

$$\pi(\theta) = \begin{cases} 1 - p_{OL} & \theta = 0 \\ \frac{p_{OL}}{\min(l_a, l_b)} & 0 < \theta \leq \min(l_a, l_b) \end{cases}$$

Instead of updating a prior for OVERLAP as in Equation 2.1, the CEPH-Genethon team updated a prior for NO OVERLAP. The mathematics are completely analogous, although the CEPH-Genethon likelihood ratio, $LOS(D)$, is the reciprocal of $L(D)$.

$$P(\theta = 0|D) = \left(1 + \left(\frac{\pi(0)}{1 - \pi(0)}\right) \cdot LOS(D)\right)^{-1} \tag{2.10}$$

$$LOS(D) = \frac{\int_{\theta>0}^{\min(l_a,l_b)} P(D|\theta)d\theta}{P(D|\theta = 0)} \tag{2.11}$$

Similar to the approach of the chromosome 16 project, the CEPH-Genethon team constructed a matrix of matches between all pairs of fragments whose relative difference was below 3 standard deviations. Let $\Omega(k)$ denote the set of all matchings between the bands of clone pair that match exactly k bands, leaving $n_a - k$ bands unmatched on the first clone and $n_b - k$ bands unmatched on the second. (This is analogous to the rightmost two sums in Equation 2.9.) Let ω denote a particular matching of bands between the clone pair.

$$P(D|\theta) = \sum_{k=0}^{\min(n_a,n_b)} \sum_{\omega \in \Omega(k)} P(\omega|\theta) \cdot P(D|\omega) \tag{2.12}$$

Exact formulae for $P(\omega|\theta)$ and $P(D|\omega)$ were not reported in the literature, though their general form was sketched in [33]. Given a matching, $P(D|\omega)$ modeled the mea-

surement error of common bands with a Gaussian distribution. The model assumed the standard deviation, σ, grew linearly with the true fragment size, x.

$$f_{x_{a_i},x_{b_j}}(x_{a_i},x_{b_j}|x) = \frac{e^{-\left[\frac{(x_{a_i}-x)^2}{\alpha x}+\frac{(x_{b_j}-x)^2}{\alpha x}\right]}}{2\pi(\alpha x)^2}$$

Poisson assumptions for restriction sites, probes, and clone ends were used to derive $P(D|\omega)$.

The same probes marked bands in the three digests, so the digests were not independent. For computational tractability, the CEPH-Genethon team treated them as independent.

2.6.3 Contig Assembly

Contigs construction occurred in the clone-clone overlap graph, where nodes represented clones and arcs represented overlap probabilities above a certain threshold of certainty. Fine ordering of contigs was not attempted. A handful of CEPH-Genethon contigs were positioned on metaphase† chromosome spreads using FISH.

2.6.4 Early Data Problems

Even in the early stages of the mapping effort, minor difficulties with the CEPH-Genethon fingerprint data were apparent. These problems included numerous chimeric clones in the YAC library (estimated at 40%), artifactual bands (at least one false positive band was found in 10% of the gels), and missing bands (the false negative rate for bands varied between 10% and 70% rate, dependent on optical density).[33] Further, the reported band measurement error, α, was suspiciously low: 0.3% for 1 kb fragments to 1.7% kb for 20 kb fragments. A 1 kb fragment, however, cannot be measured within 3 bp resolution on an agarose gel; this far exceeds the resolution of the media.[23]

Additionally, only 6 of the 10 contigs hybridized to a single location on the metaphase chromosomes during FISH verification, using pooled inter-ALU PCR†

probes from the contig clones. The remaining four hybridized to two locations, indicating chimeric clones had falsely linked noncontiguous regions of the genome.

For one such contig mapping to chromosome 1q24[4] and 10p11, its 10 constituent clones were screened individually against metaphase chromosomes using FISH. Three clones mapped to 1q24. Four mapped to 10p11. One mapped to 1q24 and 10p11. One mapped to 1q24, Xp11, and 7q36, and the last mapped to 10p11 and Xp11.

2.6.5 Comment

The exact statistics underlying the CEPH-Genethon LOS measure were not published in the literature, and the CEPH-Genethon procedure for construction was rudimentary.[5] These difficulties seem insignificant when compared to issues of data quality. Section 2.6.4 mentioned problems reported by Chantelot et al. in [8]. These and others are examined in depth in Chapter 3.

According to the CEPH-Genethon interpretation of their data, their physical map† covered between 85% and 95% of the human genome with over a thousand contigs. According to CEPH-Genethon, these contigs ranged from 2 to 10 mb in size. CEPH-Genethon did not publish their map.

2.7 The Five Projects Compared

Figure B-5 summarizes salient features of the yeast, worm, bacterium, chromosome 16, and human genome projects.

2.8 The Lander-Waterman Model

Following the yeast, worm, and bacterium projects in the mid 1980s, Lander and Waterman derived simple formulas describing how the clone library and the finger-

[4]Cytogenic† locations are denoted by chromosome, short (p) or long (q) arm, and band number from the centromere. "1q24" refers to the 24th band of the long arm of chromosome 1.

[5]It is possible that [8] was intended as an initial report, with additional data and more sophisticated analyses to follow later. However, no subsequent articles appeared in the literature.

printing scheme affect the progress of a physical mapping process.[34]

To analyze physical mapping with fingerprints, Lander and Waterman adopted a simplifying model. It considers an idealized fingerprinting method that can detect overlapping clones when they share at least a fraction θ of their length. It assumes clones are uniformly distributed across the genome. The basic model also assumes θ is constant across all clones and all clones are of constant length L.

The model uses the following variables:

G haploid† genome length in bp,

L clone length in bp,

N number of clones fingerprinted,

$\alpha = N/G$ probability per base of starting a new clone,

T minimum detectable overlap length in bp,

$c = LN/G$ redundancy of library coverage,

$\theta = T/L$, and $\sigma = 1 - \theta$.

Connected components in the clone-clone overlap graph are called apparent islands†. Islands with two or more clones are contigs.

Moving along the genome base by base, a clone begins with probability α. If no other clone begins in the next $L - T$ bases, this clone will be the last in its island. The probability this base starts a clone that ends an island is thus $\alpha(1-\alpha)^{L-T}$. This can be written as $\alpha(1 - N/G)^{(G/N)c\sigma}$, which well-approximated by $\alpha e^{-c\sigma}$ for small N/G. As there are the same number of clones that end islands as there are islands, it follows the expected islands is $G\alpha e^{-c\sigma} = Ne^{-c\sigma}$ ([34], Proposition 1.1.)

A similar argument shows the number of clones in an island is geometric with mean $e^{c\sigma}$. The probability an island contains exactly j clones is $(1 - e^{-c\sigma})^{j-1}e^{-c\sigma}$. Thus, the expected number of islands containing j clones is $N(1-e^{-c\sigma})^{j-1}e^{-c\sigma}$ (Prop. 1.2) and expected number of contigs is $Ne^{-c\sigma} - Ne^{-2c\sigma}$ (Prop. 1.2.1.) Lander and Waterman also derived the expected number of clones in island, $e^c\sigma$ (Prop. 1.3),

and the expected length of an island, $L[((e^{c\sigma} - 1)/c) + (1 - \sigma)]$ (Prop. 1.4.) By setting $\theta = 0$, any common DNA suffices for clone-clone overlap, and these results apply to undetected overlap (Prop. 1.5.) With minor modifications, the model can accommodate L and θ varying across clones (Prop. 2.)

This model allowed biologists to predict the quality of the physical map a set of experiments could be expected to generate—before conducting any experiments. Given the cost and magnitude of mapping projects, the importance of this model for strategic planning is large. Planning and progress assessment are the conventional, forward uses of this model.

Lander and Waterman note that the decreasing θ from 0.50 to 0.25 greatly speeds the progress of a mapping project, while decreasing θ from 0.25 to the theoretical limit of 0 provides relatively less improvement. They suggest θ values between 0.15 and 0.20 as sensible goals.

How efficient were these five projects in detecting overlap? The Lander-Waterman model has a less conventional, backward use: it allows one to calculate an implied θ from reported coverage and contig measures. These different performance measures are functions of θ. Solving for θ is straight-forward.

If x denotes the number of islands,

$$\theta = 1 + \frac{\ln(x/Nc)}{c}.$$

If x denotes the number of isolated clones,

$$\theta = 1 + \frac{\ln(x/N)}{2c}.$$

If x denotes the mean island size,

$$\theta = 1 - \frac{\ln(x)}{2c}.$$

If x denotes the number of contigs, then θ solves

$$x = Ne^{-(1-\theta)c} - Ne^{-2(1-\theta)c}.$$

This formula can have multiple roots in the unit interval; selecting the root closest to the value of θ produced by the other performance measures removes this ambiguity.

Figures B-5 and B-6 present θ values implied by these four performance measures for the five projects. The bacterium project achieved the lowest θ, approximately 0.2. The worm project achieved a θ of about 0.5; chromosome 16 was slightly higher. The chromosome 16 project produced far more contigs than expected, given the project's reported contig size and number of number of islands. Cloning biases or probe clustering might explain this anomaly. The yeast project achieved a θ between 0.6 and 0.7, an impressive feat considering its early date. The CEPH-Genethon project, however, performed poorly, with a θ value above 0.95.

The efficiency of the *E. Coli* project and the inefficiency of the CEPH-Genethon projects at detecting overlap reflect their respective strategies for overlap detection. The bacterium project demonstrated the power of shotgun-sequencing analysis techniques following partial digestions with multiple restriction enzymes†.

It is tempting to consider the map CEPH-Genethon might have obtained with such a strategy and a $\theta \approx 0.2$. The *E. Coli* strategy, however, would not scale up to YAC-sized inserts. Assume 200 clearly resolved bands represents an upper bound on the resolution of current gels. This limits the number of restriction sites per clone to about 19: a partial digestion of 19 sites produces about $\binom{19+2}{2} = 210$ fragments. Under the random-base model, restriction enzymes with 13 bp recognition sites are required to obtain so few restriction sites per megabase clone ($10^9/4^{12.82} \approx 19$). Restriction enzymes with such long recognition sites are rare, and it is likely that random-base model would not be a realistic representation of their occurrence along genome.

To map the human genome, CEPH-Genethon needed clones with large inserts. As the inserts were large and the resolution of gels was limited, CEPH-Genethon needed probes to select only certain bands from complete digestions†. Because of

these constraints, CEPH-Genethon could not have used the *E. Coli* approach.

The following chapter examines the quality of the CEPH-Genethon data.

Chapter 3

CEPH-Genethon Fingerprints

This chapter reviews the CEPH-Genethon fingerprint data in preparation for subsequent evaluation of pairwise overlap tests. The original Kpn-probed data doubled with the addition of a second probe, THE, in 1991. Additional clones were added to the megabase library, bringing the total to 33000 YACs. The fingerprint dataset has since stabilized; no additional experiments are planned. More recently, a dataset with YAC sizes was also released.[48]

3.1 Real Data and Simulation

Simulation provides a powerful technique to investigate the performance of a fingerprint mapping effort. Simulation can encompass any level of detail, providing greater realism than simplified analytic models (cf. [34]). Simulation is also useful when evaluating or tuning pairwise clone overlap tests, for the "right" answer is known.

A fingerprint simulator for pairwise clone overlap test evaluation would include the following: a model of clone lengths and chimeric clones; a model of clone overlap; a model of restriction site spacing; a model of probe spacing (equivalently, the number of bands per clone); a model of band measurement error; and models of false positive and false negative bands.

For example, Datta assumed constant length non-chimeric clones, overlap lengths uniformly distributed across clone lengths, the number of bands and the size of bands

drawn from empirical distributions matching the real data, Gaussian band measurement error, and Bernoulli-generated false negatives [19].

Simulation has a disadvantage: a simulator can be inaccurate. Datta found pairwise overlap statistics that performed admirably on simulated clone pairs performed less impressively on real data [19]. Clearly, his simulation differed from the real data in some unknown but substantive way. Datta's assumptions of non-chimeric and constant length clones are likely candidates, as is his assumption that THE and Kpn probes follow independent poisson processes. In reality, 40% of the clones are chimeric ([8]); clone lengths vary (Section 3.5); the two types of probes are correlated (Section 3.6); the genome consists of "probe-rich" and "probe-poor" regions ([28]); cluster or spread processes ([36]) might better describe probe locations.

This thesis eschews simulation to avoid such difficulties. The interested reader should consult Datta for a comprehensive simulation study paralleling this thesis [19].

3.2 Data Format

Figure B-7 presents a sample of the original CEPH-Genethon data format. Each clone has a seven line block of data. Line one identifies the YAC. Lines three, five and seven indicate the results of the EcoRI, PstI, and PvuII digestions, respectively. For each digestion, an integer indicates the size of a band in base pairs and is followed by a decimal number indicating the optical intensity of the band[1]. Electrophoresis sorted the bands by size.

The newer dataset differs from the earlier one in two ways. First, the three digestions appear twice, once for each probe. Second, optical density data are not included. Figure B-8 presents a sample of these data.

[1]The raw gel images were scanned, digitized, and standardized to generate these sizes. The processed band sizes and their intensities were the only data made publicly available.

3.3 Optical Densities

Figure B-9 provides a histogram of band intensity from the earlier data. The mean intensity is 0.30; the median is 0.196; the distribution has a heavy right tail. CEPH-Genethon reported that band reproducibility varied with optical intensity, with a 50% reproducibility rate for bands with optical density below 0.05.[33] 14% of observed bands had densities below this threshold.

There are two possible interpretations of these low reproducibility bands. The first interpretation declares these bands to be weak readings of true bands: false negatives. From the 50% reproducibility rate, roughly each detected weak band has a corresponding undetected band. Assuming every weak band has an undetected pair, this suggests a false negative rate of about $\frac{14\%}{86\%+14\%+14\%} = 12\%$. The second interpretation declares these bands to be spurious readings of nonexistent bands: false positives. In this case, at least 14% of the bands are false.

Such error estimates are informative, for optical intensity data were dropped from the newer release.[2] No threshold was imposed; even bands measured at optical intensity "0" in the earlier data appear in the newer dataset. Without intensity data, all bands in the newer data release appear equal, hiding a possible 12–14% false negative or false positive rate.

3.4 Band Sizes

3.4.1 Problems

Under the random-base DNA model, occurrences of a 6-cutter recognition site are well-modeled by a Bernoulli process. The distances between successive recognition sites follow a geometric distribution with mean 4^6. The placement of probes[3] may

[2]Optical intensity measurements for Kpn bands are available from the earlier dataset. Intensities for THE bands are not available. CEPH-Genethon ignored intensity data altogether in their analysis; this thesis does likewise.

[3]For all fingerprint chapters in this thesis, "probes" refers to the repetitive† elements Kpn and THE.

	EcoRI	PstI	PvuII
THE	6755	5549	6760
Kpn	7094	7691	7919

Table 3.1: Mean Band Size

be modeled with another independent Bernoulli process with a much slower rate. (The probes were purposely sparse compared to 6-cutter restriction sites, for CEPH-Genethon used rare probes to obtain a resolvable number of bands after digestion.) With these two modeling assumptions, probes may be considered "random-incidence arrivals" [21] into inter-restriction site gaps. The size of probe-containing inter-restriction gaps follows a second order Pascal distribution with mean $2 \times 4^6 = 8192$. This discrete distribution is well-approximated by its continuous counterpart, the second order Erlang.

Table 3.4.1 indicates mean band size for all three digests and both probes. That all these means fall below the random-base model prediction of 8192 is not noteworthy; DNA sequence is not Markovian. The order of these means is interesting, however. For THE, PstI had the smallest mean gap, followed by EcoRI and PvuII in an effective tie. For Kpn, the order of increasing means was EcoRI, PstI, PvuII. These rankings are not in agreement, indicating some unknown correlation between probe sites and restriction enzyme recognition sites.

Figures B-10 and B-11 present histograms and QQ-plots of band length for all six probe-digest combinations based on a random sample of 5000 bands. The QQ-plots compare the empirical distributions to second order Erlangs with matched mean. The roughly linear QQ-plots indicate the two distributions are similar. However, both informal inspection of the histograms and formal testing using the χ^2 statistic [19] indicate these distributions are not Erlang.

The coarse binning of these histogram hides data anomalies[47]. Figure B-12 presents a detailed histogram of band sizes.[4] The x axis of the histogram corresponds

[4]The six probe-digest combinations are aggregated into one. The six probe-digest pairs show an identical error structure singularly and in combination.

to band size in increments of one basepair, the given data resolution. The y axis corresponds to the number of occurrences of that band size in a random sample of 25% of the data. The figure is a line plot, with straight lines connecting adjacent nonzero counts.

One sees the quasi-Erlang structure of the length distribution as the thick black band that rises until $x \approx 8000$ and then slowly falls. The thickness of the band corresponds to the range of the counts, an indication of variance. The bottom of the thick black band remains essentially above the x axis until $x \approx 9000$.[5]

This expected quasi-Erlang structure is punctuated by a series of unexpected large spikes. Small intervals with few observations flank each spike. These nearly empty gaps are indicated by black lines reaching down to the x axis for the smaller bands. and by empty triangles beneath the larger bands. This gap-spike-gap phenomenon occurs in the same locations across all probe-digest combinations. A systematic error in the gel digitizing hardware or software is the likely explanation. The height of the spikes roughly accounts for the gaps of "missing" probability. Inter-spike spacing seems to increase exponentially with band size, suggesting an error mechanism that occurred at constant intervals along the gel. One might postulate an irregularity in a gear mechanism that drove the digitizing head across the gel films; perhaps a slight velocity hiccup on each revolution collapsed a rectangular region of the image onto a narrow bar.

Other minor anomalies afflict the length data. The dataset contains a handful of extremely long bands. The longest is 283548 bp. The genome is unlikely to contain such a large gap between 6-cutter restriction sites. The YACs average 0.8 mb in length and contain on average less than 16 probes; extrapolating this rate for the whole genome produces an overestimate of $3300 \times 16/0.8 = 66000$ probes and thus 66000 probe-containing inter-restriction site gaps. Employing a second order Erlang

[5]If x_i denotes the number occurrences of a band of size i in a sample of N bands, the joint distribution of $\{x_1, x_2, \ldots, x_{50000}\}$ is multinomial with parameters from the second order Pascal: $p_i = (i-1)(4^{-6})^2(1-4^{-6})^{i-2}$. The marginal distribution of x_i is binomial. A 95% confidence interval for x_i is $Np_i \pm 2\sqrt{Np_i(1-p_i)}$. For $N \approx 250000$, the bottom of this confidence interval hits zero near $i \approx 9000$.

with mean 6725 for the band size distribution[6], the probability that the genome contains a gap of size 283548 or larger is approximately

$$66000 \cdot \int_{283548}^{\infty} 6725^{-2} x e^{-x/6725} dx,$$

which is less than 10^{-11}. Therefore, from mathematical considerations alone, the 283548 band is highly likely to be spurious, as are another six bands longer than 250000. The optimist finds a handful of errors among 1452000 band observations encouraging. The pessimist wonders why such blatant errors were not detected and fixed, and if these errors suggest the presence of additional, undetected problems.

3.4.2 Impact

The band length distribution suffers two main problems: dramatic probability spikes and a handful of observations on the distant right tail. The remainder of the data appear reasonable.

If one assumes the probability spikes collapsed a wider observation region onto a narrow one, this anomaly only serves to reduce the resolving power of the gel over a small set of disjoint regions. As the observations that fell into these regions comprise less than 1% of the total observations, this anomaly should have little effect. If one instead assumes the spikes represent induced false positives at specific spots, this anomaly causes pairs of clones to have an extra few matching bands. As the various pairwise overlap tests are tuned to obtain selected false positive and false negative rates (Chapter 4), this should have little effect. Likewise, the excessively large bands quite rare. As no such band occurs in two clones, such bands are never matched in a pairwise clone overlap test, and thus have little effect.

The impact of these anomalies upon the pairwise overlap tests is expected to be slight. Nonetheless, such problems do erode one's confidence in the quality of these data.

[6] 6725 is the mean band size across all six probe-digest combinations.

3.5 YAC Length

Recent CEPH-Genethon data releases have included data on YAC lengths. Figure B-13 presents a histogram of lengths for the fingerprinted YACS[7]. The histogram indicates YAC length is highly variable. The mean YAC length is 910 kb, slightly longer than the earlier estimate of 810 kb by Chantelot et al.[8] A uniform distribution between 100 kb and 1750 kb provides a very crude approximation of this distribution.

3.6 Number of Bands

Figure B-14 presents histograms of the observed number of bands per YAC.

The number of bands per YAC is not poisson. This is expected, as the clones are of variable length[8]. The observed probability spikes at zero bands suggest a switching process: with probability p, the YAC has no bands, and with probability $1-p$, the YAC has a poisson number of bands with a mean proportional to its length. YACS with no bands may have come from regions of the genome lacking THE and Kpn repetitive elements or they may represent complete hybridization failure.

Figure B-15 displays the correlation between the numbers of various bands using a scatter-plot matrix. The 3x3 upper left submatrix gives plots of Kpn bands for the three digests. The points are roughly linear and indicate a high positive correlation. This is expected, as each probe should produce one band in each of the three digests. A similar pattern holds for the THE plots in the lower right 3x3 submatrix.

The lower left 3x3 submatrix plots the three digests for Kpn bands against the three digests for THE bands. The points form a diffuse cloud, but a linear correla-

[7]YACS from plates 628–989 were fingerprinted. Some of these are megabase YACS; the CEPH-Genethon megabase library consists of plates 713–996 and plates 2000+.

[8]If clone lengths were uniformly distributed between 100 kb and 1750 kb and probes followed a poisson process with rate λ probes per kb, the probability mass function for the k, the number of bands on an arbitrary YAC, would take the following form:

$$P_K(k) = \frac{1}{1.75 - 0.10} \int_{l=0.10}^{l=1.75} \frac{(\lambda l)^k e^{-\lambda l}}{k!} dl.$$

Except for the spikes at zero bands observed in the real data, this probability mass function with $\lambda = \frac{5 \text{ probes}}{0.910 \text{ mb}}$ and $\lambda = \frac{10 \text{ probes}}{0.910 \text{ mb}}$ has a similar shape to the distributions of Figure B-14.

	EcoRI Kpn	PstI Kpn	PvuII Kpn	EcoRI THE	PstI THE	PvuII THE
EcoRI Kpn	1					
PstI Kpn	0.93	1				
PvuII Kpn	0.94	0.93	1			
EcoRI THE	0.62	0.62	0.62	1		
PstI THE	0.62	0.63	0.62	0.94	1	
PvuII THE	0.63	0.62	0.63	0.94	0.95	1

Table 3.2: Correlation Coefficients, Number of Bands

tion is still seen. Table 3.6 provides all pairs of correlation coefficients between these counts. The THE-Kpn correlation coefficients exceed 0.6, indicating these two repetitive elements frequently occur together on the genome. This has two implications. The first is that doubling the data by adding the THE probe did not produce as much additional coverage as might have been obtained with an independent or, better yet, negatively correlated probe. The second implication is that the regions of the genome covered by Kpn-base contigs in 1992 should have stronger overlap results from the additional THE probes in 1993.

3.7 Band Measurement Uncertainty

As mentioned in Section 2.6.4, CEPH-Genethon reported a dubiously low standard deviation for band measurement error: 0.3% for 1 kb fragments to 1.7% kb for 20 kb fragments.

To estimate this rate de novo from the data, a small number of plates in the YAC library with highly similar fingerprints in adjacent wells were identified.[9] Plate contamination is the most likely explanation of this phenomenon. This provides repeated measurements of (what is highly likely to be) the same band. The standard deviation, σ, was observed to vary slowly with band length, x. Point estimates of $\sigma(x)$ were computed for $x = 1, 2, \ldots$ kb using overlapping 2 kb windows. A quadratic

[9]The measure was the $\frac{2s}{a+b}$ test described in Section 4.1 with a threshold of 0.95 on YACS with 40 or more bands.

curve was fit to these estimates using least squares regression ($R^2 > 0.95$):

$$\sigma(x) = 41.9 - 0.0005x + 2.7 \times 10^{-7}x^2. \tag{3.1}$$

The details of this estimation are provided by Datta[19]. This thesis employs Equation 3.1 to model band measurement error.

The following chapter considers four pairwise overlap tests.

Chapter 4

Pairwise Fingerprint Tests

This chapter compares the performance of five pairwise clone overlap tests using the CEPH-Genethon fingerprint data. It presents the underlying models motivating the tests and describes their possible uses. The five tests are named `Trinomial`, `Match`, `Entropy`, `KPN`, and `THE`. The first three were developed at MIT for this thesis; the last two were developed at CEPH-Genethon for their human genome mapping effort [8].

Formally, each test is an indicator function H with parameter $\vec{\chi}$ that decides if the fingerprint data from two clones, $\vec{d_1}$ and $\vec{d_2}$, are sufficiently similar to indicate overlap:

$$H_{\vec{\chi}} : (\vec{d_1}, \vec{d_2}) \rightarrow \{0, 1\}.$$

Test performance is measured using false positive and false negative error rates.

$$f_p(H_{\vec{\chi}}(\vec{d_1}, \vec{d_2})) = P(H_{\vec{\chi}} = 1 | \text{ clone 1 and clone 2 do not overlap}) \qquad (4.1)$$

$$f_n(H_{\vec{\chi}}(\vec{d_1}, \vec{d_2})) = P(H_{\vec{\chi}} = 0 | \text{ clone 1 and clone 2 do overlap}) \qquad (4.2)$$

The parameter vector, $\vec{\chi}$, represents all the constants involved in the test, including the test's threshold value upon which the overlap decision is based. Note the false positive and false negative rates depend on $\vec{\chi}$. Varying the test's decision threshold changes these error rates.

Each test assumes a different model of the fingerprint data. Stronger assumptions

lead to simpler models and, perhaps, to weaker tests. The tests may be ranked by complexity. The simplest is `Trinomial`; next follows `Match` and `Entropy`; `KPN` and `THE` are the most involved.

Sections 4.1 through 4.4 describe these five tests. Section 4.5 explains the method used to evaluate them and Section 4.6 presents results. Section 4.7 concludes the chapter discussing how the tests might be used.

4.1 The `Trinomial` Test

`Trinomial` employs the simplest model of the fingerprint data. The test uses a maximum likelihood estimator [35] for θ, the fraction of overlap between the two clones, and declares overlap if θ is large enough.

Assumptions of `Trinomial` model

1. All clones are the same length. Distance is rescaled so this length is 1.
2. Band placement follows a homogeneous poisson process.
3. Every band may be assigned one of three designations:
   (a) belonging only to clone 1,
   (b) belonging only to clone 2,
   (c) or belonging to both clone 1 and clone 2.

   No errors are made in these assignments.
4. There are no false positive bands: every band is real.
5. There are no false negative bands: no bands are lost.
6. The six probe-clone digests are independent.

7. Clones are not chimeric.[1]

8. All band lengths are equally likely across the gel.

Let $\theta \in [0,1]$ denote the length of the region shared by clone 1 and clone 2. For each of the $i = 1 \ldots 6$ digests[2], let X_i denote bands unique to clone 1, Y_i denote bands unique to clone 2, and S_i denote bands shared by both clones. Let $N_i = X_i + S_i + Y_i$ denote the total number of bands in each digest.

From assumption 3. the two clones overlap if any $S_i > 0$. In reality, however, the S_i will involve errors, so $\exists i : S_i > 0$ is not a useful test statistic. Instead, `Trinomial` considers the maximum likelihood estimator of θ. If θ_{MLE} is large enough, `Trinomial` declares overlap.

Given θ, the probability of observing the matched and unmatched bands follows a trinomial distribution.

$$P_\theta(\vec{X}, \vec{S}, \vec{Y} | \vec{N}) = \prod_{i=1}^{6} \frac{n_i!}{x_i! s_i! y_i!} \left(\frac{1-\theta}{2-\theta}\right)^{x_i} \left(\frac{\theta}{2-\theta}\right)^{s_i} \left(\frac{1-\theta}{2-\theta}\right)^{y_i} \tag{4.3}$$

This is written as a likelihood function of θ, $L_{\vec{x},\vec{s},\vec{y}}(\theta)$,

$$L_{\vec{x},\vec{s},\vec{y}}(\theta) = k \cdot \left(\frac{\theta}{2-\theta}\right)^{\sum s_i} \left(\frac{1-\theta}{2-\theta}\right)^{(\sum x_i + \sum y_i)} \tag{4.4}$$

differentiated,

$$\frac{\partial L_{\vec{x},\vec{s},\vec{y}}(\theta)}{\partial \theta} =$$

$$-k \cdot \frac{\theta \sum x_i + \theta \sum y_i - 2\sum s_i + 2\theta \sum s_i}{\theta(\theta-1)(\theta-2)} \left(\frac{\theta}{2-\theta}\right)^{\sum s_i} \left(\frac{1-\theta}{2-\theta}\right)^{(\sum x_i + \sum y_i)}$$

[1]None of these five tests explicitly model chimeric clones, for chimerism only serves to make the overlap region smaller. Chimeras pose more difficulty for contig assembly algorithms than for pairwise overlap algorithms.

[2]As `Trinomial`, `Match`, and `Entropy` all assume the six probe-digest combinations are independent, "probe-digest combination" is shorted to "digest" with no lack of accuracy.

and solved for zero, yielding

$$\theta_{MLE} = \frac{2\sum s_i}{\sum x_i + \sum y_i + 2\sum s_i}. \tag{4.5}$$

The second derivative is negative, $\frac{\partial^2 L_{x,s,y}(\theta_{MLE})}{\partial \theta^2} < 0$, indicating θ_{MLE} maximizes Equation 4.4.

Equation 4.5 has an intuitive explanation. Suppose one wished to estimate θ but could only observe $\sum X_i$ and $\sum S_i$. In this case, the best estimate for θ would be $\theta_1 = \sum s_i/(\sum s_i + \sum x_i)$. Alternatively, if one wished to estimate θ but could only observe $\sum Y_i$ and $\sum S_i$, the best estimate for θ would be $\theta_2 = \sum s_i/(\sum s_i + \sum y_i)$. These estimates, weighted according to proportion of the data they represent, also produce θ_{MLE}.

$$\theta_1 \left(\frac{\sum x_i + \sum s_i}{\sum x_i + \sum y_i + 2\sum s_i} \right) + \theta_2 \left(\frac{\sum y_i + \sum s_i}{\sum x_i + \sum y_i + 2\sum s_i} \right) = \frac{2\sum s_i}{\sum x_i + \sum y_i + 2\sum s_i}$$

The implementation of `Trinomial` is particularly simple. For each of the six digests, Equation 3.1 is used to match bands within 3 standard deviations of their midpoint. Matching is done in a greedy, nearest neighbor fashion. The bands are ordered so this greedy approach yields the most matchings possible. This computes x_i, s_i, and y_i for each lane. Equation 4.5 then produces θ_{MLE}. If this is large enough, $\theta_{MLE} \geq \theta_{CRIT}$, `Trinomial` declares the clones overlapping.

The performance of this test on real data for various settings of θ_{CRIT} are discussed in Section 4.6.

4.2 The `Match` Test

`Trinomial` does not acknowledge that the length of clones differs widely and that errors are made in band matching. The second test, `Match`, adds these features. Like `Trinomial`, `Match` is based on a maximum likelihood estimator of the length of the overlap region common to both clones.

Assumptions of Match model

1. The length of the two clones, L_1 and L_2, are known.
2. Band placement follows a homogeneous poisson process.
3. Every band may be assigned one of three designations:
   (a) belonging only to clone 1,
   (b) belonging only to clone 2,
   (c) or belonging to both clone 1 and clone 2.
4. False positive matching errors (two bands are matched which are not the same) and false negative matching errors (two bands that are the same are not matched) occur as described below.
5. There are no false positive bands: every band is real.
6. There are no false negative bands: no bands are lost.
7. The six probe-clone digests are independent.
8. Clones are not chimeric.
9. All band lengths are equally likely across the gel.

Let $\theta \in [0, \min(L_1, L_2)]$ denote the length of the region shared by the two clones. For each of the $i = 1 \ldots 6$ digests, let X_i denote bands truly unique to clone 1, Y_i denote bands truly unique to clone 2, and S_i denote bands truly shared by both clones. From assumption 2, these follow poisson distributions with rate λ probes per kb.

$$P_{X_i}(x_i) = \frac{(\lambda(L_1 - \theta))^{x_i} e^{-\lambda(L_1-\theta)}}{x_i!}$$

$$P_{Y_i}(y_i) = \frac{(\lambda(L_2 - \theta))^{y_i} e^{-\lambda(L_2-\theta)}}{y_i!}$$

$$P_{S_i}(s_i) = \frac{(\lambda\theta)^{s_i} e^{-\lambda\theta}}{s_i!}$$

`Match` assumes that missed matches follow a Bernoulli process[21] with probability p_{LOSE}. Let R_i denote the number of matched band pairs that were declared non-matching, due to a false negative mismatch error.

$$P_{R_i|S_i}(r_i|s_i) = \binom{s_i}{r_i} p_{LOSE}{}^{r_i}(1 - p_{LOSE})^{(s_i - r_i)}$$

Several values were tried for the rate of this Bernoulli loss process; $p_{LOSE} = 0.10$ produced the best empirical results.

To model false matches, `Match` adopts the same methodology as the worm project's *PROBCOINC*, as described in Section 2.3.2. Note Equation 2.5 and Equation 4.6 below assume all band lengths are equally likely. This assumption was empirically justified for the worm project (see Section 2.3.1). `Match` simply assumes this is true.[3] Let Q_i denote the number of incorrect matches.

$$P_{Q_i|X_i,Y_i}(q_i|x_i, y_i) =$$

$$\binom{\min(x_i,y_i)}{q_i}(1 - (1 - p_{gain})^{(\max(x_i,y_i))})^{q_i}(1 - (1 - (1 - p_{gain})^{(\max(x_i,y_i))}))^{(\min(x_i,y_i)-q_i)} \tag{4.6}$$

The model implied by Equation 4.6 was justified by empirical data and linear regression. The approximate expected value of Q_i,

$$\begin{aligned} E(Q_i) &= \min(x_i, y_i)(1 - (1 - p_{gain})^{(\max(x_i,y_i))}) && (4.7)\\ &\approx p_{gain}(\max(x_i, y_i))(\min(x_i, y_i)) && (4.8)\\ &= x_i \cdot y_i \cdot p_{gain}, && (4.9) \end{aligned}$$

indicates that q_i should vary linearly in the product $x_i y_i$. To validate Equation 4.9, 500 random pairs of clones were generated. The number of matching bands in each pair was determined using the methods described in `Trinomial`.[4] These (x_i, y_i, q_i) data were well-fit with a linear model with zero intercept, with an R^2 exceeding

[3] Assumption 9.

[4] As discussed in Section 4.5.1, almost none of these pairs should consist of clones that actually overlap on the genome. As a result, almost all of the matches are false, providing data to estimate p_{gain}.

0.95. Figure B-16 plots real against fitted q_i data, demonstrating the quality of this model. Two methods were used to estimate p_{gain}: the slope of the regression line, and numerical maximum likelihood estimation on the joint density of all the (x_i, y_i, q_i) tuples via Equation 4.6. Both produced similar results: $p_{gain} \approx 0.0065$.

Let A_i denote the number of bands that appear to be unique to clone 1, B_i denote the number of bands that appear to belong to both clones, and C_i denote the number of bands that appear to be unique to clone 2.

$$A_i = X_i - Q_i + R_i \tag{4.10}$$

$$B_i = S_i + Q_i - R_i \tag{4.11}$$

$$C_i = Y_i - Q_i + R_i \tag{4.12}$$

Clearly, A_i, B_i, C_i, X_i, Y_i, S_i, Q_i, and R_i must remain nonnegative.

Analogously to Equation 4.3, the probability of observing the data $(\vec{a}, \vec{b}, \vec{c})$ for a given θ is

$$P_\theta(\vec{A}, \vec{B}, \vec{S}) = \prod_{i=1}^{6} \sum_{\substack{x_i,y_i,s_i,q_i,r_i \geq 0 \\ \text{such that} \\ x_i - q_i + r_i = a_i \\ s_i + q_i - r_i = b_i \\ y_i - q_i + r_i = c_i}} P(q_i|x_i, y_i) \cdot P(r_i|s_i) \cdot P_\theta(x_i) \cdot P_\theta(y_i) \cdot P_\theta(s_i). \tag{4.13}$$

Equations 4.10 through 4.12 allow this sum to be written efficiently via auxiliary functions, $h_j(q_i, r_i, a_i, b_i, c_i, \theta)$.

$$h_1(q_i, r_i, a_i, b_i, c_i, \theta) = P(q_i|x_i = a_i + q_i - r_i, y_i = c_i + q_i - r_i) \tag{4.14}$$

$$h_2(q_i, r_i, a_i, b_i, c_i, \theta) = P(r_i|s_i = b_i - q_i + r_i) \tag{4.15}$$

$$h_3(q_i, r_i, a_i, b_i, c_i, \theta) = P_\theta(x_i = a_i + q_i - r_i) \tag{4.16}$$

$$h_4(q_i, r_i, a_i, b_i, c_i, \theta) = P_\theta(y_i = c_i + q_i - r_i) \tag{4.17}$$

$$h_5(q_i, r_i, a_i, b_i, c_i, \theta) = P_\theta(s_i = b_i - q_i + r_i) \tag{4.18}$$

With these auxiliary functions, Equation 4.13 may be written as a likelihood function for θ.

$$L_{\vec{a},\vec{b},\vec{c}}(\theta) = \prod_{i=1}^{6} \sum_{q_i} \sum_{r_i} \prod_{j=1}^{5} h_j(q_i, r_i, a_i, b_i, c_i, \theta) \tag{4.19}$$

Like `Trinomial`, Equation 3.1 is used to match bands within 3 standard deviations of their midpoint for each digest in a greedy nearest neighbor fashion. This yields x_i, s_i, and y_i for each lane. Unlike the MLE used in `Trinomial`, there is no simple analytical formula for the value of θ that maximizes Equation 4.19. Nonetheless, an approximate θ_{MLE} can be found numerically. `Match` substitutes $\lceil \frac{\min(L_1, L2)}{100 \text{ kb}} \rceil$ different values for θ, spaced uniformly across the interval $[0, \min(L_1, L_2)]$, into Equation 4.19. `Match` then takes the θ producing the highest likelihood as θ_{MLE}. Evaluating the likelihood at a handful of points is a crude optimization technique. Nonetheless, as doubling the density of search points did not increase the efficiency of `Match`, it appears sufficient.

The performance of `Match` on real data for various settings of θ_{CRIT} are discussed in Section 4.6.

4.3 The Entropy Test

As discussed in Section 3.4.1, CEPH-Genethon fingerprint bands are not uniformly distributed across the gels. Rather, the distribution is bell-shaped with a long right tail. `Trinomial` and `Match` give equal credit to all matchings. By chance alone, the clustering of band lengths near the mean will produce spurious matches. `Entropy` remedies this deficiency by using the empirical distributions from Figures B-10 and B-11 to model band length.

Unlike the MLE-based `Trinomial` and `Match`, `Entropy` is based on the theory of statistical entropy [57]. Consider a random variable ψ that has the true distribution $p_{\psi|H_0}(\psi|H_0)$ under the null hypothesis, H_0. Let $p^*(\psi)$ denote an empirical distribution of observations of ψ. The entropy of these observations is defined as the expected

value of the natural log of their probabilities.

$$\mathcal{K} \equiv E(-\log[p^*(\psi)]) = -\int \log[p^*(\psi)] p_{\psi|H_0}(\psi|H_0) d\psi \tag{4.20}$$

If $\mathcal{K}$ is small, the empirical observations $p^*(\psi)$ are consistent with H_0, and the data do not contradict the null hypothesis. If $\mathcal{K}$ is too large, the null hypothesis must be rejected.

`Entropy` takes this approach. Indicator random variables denoting the state of each band (matched or not matched) play the role of ψ in Equation 4.20. The null hypothesis is

$$H_0: \quad \text{Clones 1 and 2 do not overlap.}$$

Intuitively, `Entropy` assigns a score for matching bands. Rarer bands receive a higher score. If the total score is too high, indicating too much matching, `Entropy` rejects the null hypothesis and declares the clones to overlap.

Assumptions of `Entropy` model

1. Band placement follows a homogeneous poisson process.
2. Every band may be assigned one of three designations:
   (a) belonging only to clone 1,
   (b) belonging only to clone 2,
   (c) or belonging to both clone 1 and clone 2.

   No errors are made in these assignments.
3. There are no false positive bands: every band is real.
4. There are no false negative bands: no bands are lost.
5. The six probe-clone digests are independent.
6. Clones are not chimeric.

7. Band lengths follow the empirical distributions of Figures B-10 and B-11.

The notation from Section 2.2.2 is again used. For each of the six digests, let "reference clone" denote the clone with more bands, and "comparison clone" denote the clone with fewer bands. Considering each digest in turn, let $x_{a_1}, x_{a_2}, \ldots, x_{a_n}$ denote the sizes of the a_n bands in the reference clone and $x_{b_1}, x_{b_2}, \ldots, x_{b_n}$ denote the sizes of the b_n bands in the comparison clone. Let $m_{b_1}, m_{b_2}, \ldots, m_{b_n}$ indicate paired bands in the comparison clone: $m_{b_i} = 1$ if comparison band i matches some band in the reference clone.

Under the null hypothesis, the distribution of each match-indicating Bernoulli random variable is known. Consider an arbitrary m_i, corresponding to a band of size x_{b_i}. Let p_{one} denote the chance a single band from the reference clone happens to be close enough to match this band. Equation 3.1 defines $\sigma(x)$, the band measurement error. For each digestion, band length is assumed to follow a second order Erlang distribution with mean $\overline{x}$. As determined by empirical experimentation, `Entropy` worked best with a four standard deviation window for matching.

$$p_{one} = \int_{x_{b_i}-2\sigma(x_{b_i})}^{x_{b_i}+2\sigma(x_{b_i})} (2/\overline{x})^2 x e^{-(2/\overline{x})x} dx$$

Let p_{any} denote the chance any band in the reference clone lands close enough to band x_{b_i} to match it.

$$p_{any} = 1 - (1 - p_{one})^{a_n}$$

The mass function for the indicator random variable, m_i, follows immediately,

$$P_{M_i|H_0}(m_i|H_0) = \begin{cases} p_{any} & m_i = 1 \\ 1 - p_{any} & m_i = 0, \end{cases}$$

and the function $\mathcal{K}(m_i)$ provides the entropy of the m_i.

$$\mathcal{K}(m_i) = \begin{cases} -\log(p_{any}), & m_i = 1 \\ -\log(1 - p_{any}), & m_i = 0 \end{cases}$$

The mean and variance of $\mathcal{K}(m_i)$ are straightforward.

$$E(\mathcal{K}(m_i)) = (-\log(p_{any}))p_{any} + (-\log(1-p_{any}))(1-p_{any})$$

$$\sigma^2(\mathcal{K}(m_i)) =$$
$$((-\log(p_{any}))^2 p_{any} + (-\log(1-p_{any}))^2(1-p_{any}))$$
$$-((-\log(p_{any}))p_{any} + (-\log(1-p_{any}))(1-p_{any}))^2$$

`Entropy` sums the entropies of each indicator random variable. Though the m_i are not iid[5], they do possess finite moments, and the Central Limit Theorem indicates their sum converges in distribution to a Gaussian. For a test statistic, `Entropy` normalizes their sum. The sums in Equation 4.21 are taken over all digests and all bands.

$$\mathcal{K} = \frac{\sum \mathcal{K}(m_i) - \sum E[\mathcal{K}(m_i)]}{\sqrt{\sum \sigma^2[\mathcal{K}(m_i)]}} \mapsto Z(0,1). \tag{4.21}$$

`Entropy` uses the same approach as `Trinomial` and `Match` to match bands. `Entropy` then computes $\mathcal{K}$ with Equation 4.21. If $\mathcal{K}$ is sufficiently large, $\mathcal{K} > \mathcal{K}_{CRIT}$, `Entropy` rejects the null hypothesis and declares the clones to overlap.

The performance of `Entropy` on real data for various settings of $\mathcal{K}_{CRIT}$ are discussed in Section 4.6.

4.4 The KPN and THE Tests

The last two statistics, `KPN` and `THE`, are somewhat "black boxes." CEPH-Genethon has only sketched their form in the literature. Section 2.6.2 presents this material. Lacking the tests, `THE` and `THE` are evaluated through their performance on two samples of clone pairs obtained from CEPH-Genethon[31]. The first sample consists of 4500 random pairs of clones. The second sample consists of 3000 STS-linked pairs of clones[6].

The `THE` and `KPN` statistics could be combined to produce hybrid tests:

[5]Indeed, assigning different scores for matching different sized bands is the core idea of `Entropy`.
[6]Test evaluation is discussed in Section 4.5.

1. An overlap is significant if THE $> T_1$ and KPN $> T_2$.

2. An overlap is significant if THE $> T_1$ or KPN $> T_2$.

If the tests were independent, the false positive false negative rates for these hybrid tests could be determined directly from the error rates of the two individual tests.

$$f_p(\texttt{THE} > T_1 \text{ and } \texttt{KPN} > T_2) = f_p(\texttt{THE} > T_1) \cdot f_p(\texttt{KPN} > T_2)$$

$$f_n(\texttt{THE} > T_1 \text{ and } \texttt{KPN} > T_2) = 1 - (1 - f_n(\texttt{THE} > T_1))(1 - f_n(\texttt{KPN} > T_2))$$

$$f_p(\texttt{THE} > T_1 \text{ or } \texttt{KPN} > T_2) = 1 - (1 - f_p(\texttt{THE} > T_1))(1 - f_p(\texttt{KPN} > T_2))$$

$$f_n(\texttt{THE} > T_1 \text{ or } \texttt{KPN} > T_2) = f_n(\texttt{THE} > T_1) \cdot f_n(\texttt{KPN} > T_2)$$

As one would expect given the correlation of the Kpn and THE probes, THE and KPN are linearly dependent, having a correlation of 0.65. Figure B-17 plots KPN against THE scores.

This thesis does not consider the hybrid tests. The individual performance of KPN and THE on real data for various settings of T_1 and T_2 are discussed in Section 4.6.

4.5 Evaluating the Tests

A test bed of clones was assembled to evaluate the pairwise fingerprint overlap tests. Two sets of clone pairs were obtained from CEPH-Genethon with THE and KPN scores attached. One set consisted of random pairs of clones, the other of STS singly-linked pairs. These sets were supplemented with additional random and STS singly-linked pairs. Both MIT and CEPH-Genethon unambiguous† STS content data were used for these single linkages. Deficient clone pairs—pairs with a clone lacking length or fingerprint data—were deleted. The final test bed consists of 16109 clone pairs. Each clone pair is either random, R, or singly-linked, S. Each clone pair has CEPH-Genethon test results, T, or does not, N. Table 4.5 indicates the breakdown of clone pairs across these types. R outnumber S pairs, for the generation of S pairs is limited by availability of unambiguous STS data. The surfeit of R pairs is helpful, however,

16109 total pairs	11185 R pairs	4924 S pairs
6766 T pairs	4057 R T pairs	2709 S T pairs
9343 N pairs	7128 R N pairs	2215 S N pairs

Table 4.1: Breakdown of Test bed clone pairs

for this thesis emphasizes the false positive error rate $f_p(H_{\vec{\chi}})$, which is determined by test performance on R pairs.

4.5.1 Chance Matches

With over 11000 R pairs, there is the concern that some of these pairs, simply by chance, might actually overlap on the genome. Such "R" pairs would be S pairs in disguise, and increase the false positive rates of all the tests.

The following simple model indicates this concern is unfounded. Possibly one, and at most two, of the 11000 R pairs should actually overlap. Let L denote the clone length, here assumed to be constant. Let p_{chim} denote the probability a clone is chimeric, modeled by two disjoint $L/2$ intervals. Let G denote the length of the genome and θ denote the length of the minimum common region needed to detect overlap. Assume the clones are uniformly distributed across the genome.

The probability of the two clones in a R pair overlapping by chance decomposes

into three cases.

$$P(\text{clones match by chance}) =$$
$$P(\text{clones match by chance}|\text{neither clone chimeric})P(\text{neither clone chimeric})+$$
$$P(\text{clones match by chance}|\text{one clone chimeric})P(\text{one clone chimeric})+$$
$$P(\text{clones match by chance}|\text{both clones chimeric})P(\text{both clones chimeric}) \tag{4.22}$$

The probabilities of the conditioning events are straight-forward.

$$P(\text{neither clone chimeric}) = (1 - p_{chim})^2 \tag{4.23}$$

$$P(\text{one clone chimeric}) = 2(p_{chim})(1 - p_{chim}) \tag{4.24}$$

$$P(\text{both clones chimeric}) = (p_{chim})^2 \tag{4.25}$$

Consider two non-chimeric clones, representing the intervals $[x_1, x_1 + L]$ and $[x_2, x_2 + L]$ on the genome. If $x_1 - (L - \theta) < x_2 < x_1$, or if $x_1 < x_2 < x_1 + (L - \theta)$, the two clones overlap. Given the assumption of uniformly distributed clones, $P(\text{clones match by chance}|\text{neither clone chimeric})$ follows. Similar geometric arguments produce the other two conditional probabilities.

$$P(\text{clones match by chance}|\text{neither clone chimeric}) = 1 - \left(1 - \frac{2L - 2\theta}{G}\right) \tag{4.26}$$

$$P(\text{clones match by chance}|\text{one clone chimeric}) = 1 - \left(1 - \frac{\frac{3}{2}L - 2\theta}{G}\right)^2 \tag{4.27}$$

$$P(\text{clones match by chance}|\text{both clones chimeric}) = 1 - \left(1 - \frac{L - 2\theta}{G}\right)^4 \tag{4.28}$$

Figure B-18 presents Equation 4.22 as a function of θ for parameter values representative of the human genome and the CEPH-Genethon clone library[7]. The Lander-Waterman calculations of Section 2.8 indicate the minimal detectable overlap for the CEPH-Genethon data using the CEPH-Genethon test exceeds 90% of the clone

[7] $L = 0.9$ mb, $p_{chim} = 0.4$, $G = 3300$, and $\theta \in [0 \ldots 0.9]$ mb.

length. For $L = 0.9$ mb, $\theta = 0.8$ mb, and Figure B-18 indicates less than one match-by-chance is expected in 10000 R pairs.

As a result, the R pairs provide a sample of non-overlapping clones. The S pairs provide a sample of overlapping pairs. As discussed in Section 2.5.1, random incidence arguments indicate that STS-containing overlaps are larger than average.[8]

4.5.2 Using the Test Bed

The Trinomial, Match, and Entropy statistics were computed for each of the 16109 clone pairs. This computation took about 100 minutes on a Sparc Sun workstation; the MLE search in Match consumed most of this time. For each test, the pairs were sorted by the test value. (For KPN and THE, N pairs were omitted.) By scanning the pairs in sorted order and counting R and S pairs, empirical estimates of $f_p(H_{\vec{\chi}})$ and $f_n(H_{\vec{\chi}})$ were obtained for each threshold value for each test. Figures B-19, B-20, B-21, B-22, and B-23 display false negative rates (indicated with a "+" symbol) and false positive rates (indicated with a "$\diamond$" symbol) for each test as a function of threshold. Figures B-24, B-25, B-26, B-27, and B-28 plot the false positive against the false negative rate for each test for each threshold, presenting the efficiency of each test. Discontinuities in Figure B-20 reflect the grid search MLE maximization in Match.

4.6 Test Results

Figure B-29 presents the five efficiency plots overlaid for comparison. Entropy is the most powerful test over a broad regime of false positive rates, for it achieves the lowest false negative rate. THE is the next most powerful test, followed by Trinomial. One interesting observation is that different tests perform better in different regimes. Match, for example, outperforms Trinomial for false positive rates exceeding 1/2. Figure B-30 presents an enlarged view of Figure B-29. For small false positive rates, THE is most efficient, followed by KPN. Table 4.6 presents the tests' false negative rates

[8]For the case of non-chimeric overlapping clones, the overlap region is uniform on $[0, L]$ and has mean length $L/2$. Conditioning on a STS in the overlap region increases the mean length to $2L/3$.

	$f_p = \frac{1}{50}$	$f_p = \frac{1}{100}$	$f_p = \frac{1}{500}$	$f_p = \frac{1}{1000}$	$f_p = \frac{1}{5000}$
THE	0.39	0.42	0.48	0.51	0.64
KPN	0.49	0.53	0.67	0.63	0.70
Entropy	0.39	0.43	0.53	0.57	0.80
Trinomial	0.59	0.60	0.70	0.75	0.88
Match	0.84	0.92	0.98	0.99	1.00

Table 4.2: False Negative Rates

as function of false positive rate.

4.7 Uses of Pairwise Overlap Tests

The utility of a statistical test must be evaluated in context. To conclude these chapters on CEPH-Genethon fingerprint data and pairwise fingerprint overlap tests, possible uses of these data and tests are discussed. This section draws upon MIT experiments and analysis not treated in this thesis. The interested reader should also consult Data ([19]), who considers the following material in much greater depth.

Ordered Contigs

The CEPH-Genethon fingerprint data are not strong enough to order clones within unordered contigs. Two experiments support this claim. First, both manual and automated restriction mapping efforts using the CEPH-Genethon data failed, even when the clone ordering was known from reliable STS data. Second, `Entropy` was unable to order a 50 clone contig from the MIT chromosome 22 mapping effort. `Entropy` fractured the contig into many small pieces with spurious links between them.[9]

The problem, of course, could be `Entropy`, not the CEPH-Genethon data. More likely, the blame rests on the poor quality of fingerprint data set. These data were unable to support genomic mapping in 1991. These data, even when employed by the

[9] `Entropy`'s overlap threshold was set to produce $f_p \approx 0.001$, $f_n \approx 0.4$.

more sophisticated CEPH-Genethon tests, performed poorly on the pairwise overlap problem, as shown in Table 4.6. It is impossible to give formal proof that these data could not support mapping, but the failure of reasonable approaches indicate that this is most likely the case.

Unordered Contigs

The CEPH-Genethon library consists of 33000 clones, or 5.5×10^8 clone pairs. With five-fold genomic coverage, each clone overlaps about ten others, ignoring the difficulties created by chimeras. Suppose `Entropy` was able to detect small overlaps (though from all appearances, it cannot), even at the $f_p = 1/5000$ threshold. Of the 330000 pairs of overlapping clones, `Entropy` would detect 40%, or 130000. Of the $(\binom{33000}{2} - 10 \times 33000)$ pairs of non-overlapping clones, `Entropy` would incorrectly detect 1/5000, or 110000. With the same magnitude of true and false detects, overlaps declared by `Entropy` are about as likely false as true.

For this reason, `Entropy` does not appear capable of producing valid unordered contigs by comparing all pairs of clones in the library. The CEPH-Genethon test also failed at this task in 1992. Again, poor quality data is the likely culprit.

Single Linkage

The CEPH-Genethon data can assist with the single linkage† problem. Given two STSs with only one clone in common, it is not clear whether the STSs are close on the genome or whether the linking clone is chimeric. Let A denote the set of clones hit by STS 1 and B denote the set of clones hit by STS 2. Single linkage implies $|A \cap B| = 1$. As $|A|$ and $|B|$ are usually small (on average, each STS hits fewer than 8 clones), there are only a small number of pairs $(i, j) : i \in A, j \in B$. A pairwise fingerprint overlap test could examine all such pairs. If enough pairs appeared to overlap, the two STSs are likely to be close on the genome.

Disambiguating STS Addresses

STS content is determined using pools of clones, and experimental error generates incomplete or ambiguous† addresses. `Entropy` has been used successfully to resolve some of these cases by finding overlap between definite and ambiguous clones. In the laboratory, this technique has resolved roughly half of ambiguous addresses. The error rate of these resolved addresses is less than 5%, the unpooled YAC-STS false negative rate.[28]

4.8 Conclusion

The CEPH-Genethon fingerprint data are useful for small tasks. The CEPH-Genethon fingerprint data are of insufficient quality to support genomic mapping, the very purpose for which they were generated. More sophisticated statistics such as `KPN` and `THE` outperform simpler statistics for the pairwise overlap problem, especially when a low false positive rate is required. Of the simpler tests, `Entropy` performs surprisingly well. The overall lackluster performance of these tests is most likely due to poor data quality.

Chapter 5

Mapping with ALU-PCR Probes

This chapter examines the CEPH-Genethon ALU-PCR† mapping effort. Unlike the fingerprint-based mapping strategies described in previous chapters, the CEPH-Genethon ALU-PCR approach uses probes to identify overlapping clones. Section 5.1 reviews the probe-based mapping literature. Section 5.2 introduces the CEPH-Genethon ALU-PCR mapping project. Section 5.3 highlights problems with the resulting map and Section 5.4 attempts to remedy them.

5.1 Probe Mapping Literature Review

Some mapping efforts employ "single-copy" probes, probes occurring only once on the genome. Others use "multiple-copy" probes, probes with several copies scattered across the genome. The distinction is important for mapping algorithms.

5.1.1 Mapping with Single-Copy Probes

With perfect data, non-chimeric clones, and single-copy probes, the problem of probe-ordering reduces to the consecutive ones matrix problem. This problem may be solved in linear time using P-Q trees [12]. Unfortunately, this approach does not generalize to imperfect data or chimeric clones [26]. From the perspective of worst-case compu-

tational complexity, realistic probe-based mapping problems are NP-complete [25].[1] Nonetheless, given data of sufficient quantity and quality, heuristics perform well on real instances of these problems.[2]

Early projects ordered probes without formal optimization methods, relying instead on expert judgment to produce reasonable-looking orders ([24], [49]). To order probes across the human Y chromosome, Foote et al. solved a "noisy" consecutive ones problem by eye. They manipulated their STS-YAC† incidence matrix manually with a microcomputer spreadsheet program to generate their map. [23].

Mott et al. adopted a TSP-based optimization probe-ordering strategy for the *S. Pombe* project [38]. Let C_i denote the set of clones hit by STS i. Mott et al. defined the distance between STSs i and j as the fraction of clones hit by only one of the two STSs: $d(i,j) = (|C_i \cup C_j| - |C_i \cap C_j)/|C_i \cup C_j|$.[2] If the STSs in the permutation are labeled $(s_1, s_2, \ldots, s_n)$, the objective function takes the form $\mathcal{C} = \sum_{i=1}^{n-1} d(i, i+1)$. Using two-opt and simulated annealing [29], the *S. Pombe* team sought the permutation with the shortest TSP path. This TSP-style objective function for probe ordering has been strongly criticized by other researchers ([2], [22]).

The CEPH-Genethon Chromosome 21 project ([13], [14]) also used simulated annealing to order STS probes. Two-opt, three-opt, single probe shift, and single probe swap operators defined the neighborhood of a permutation. Simulated annealing was used to minimize the total sum of the gaps across the clones.[46] Given a permutation with STSs labeled $(s_1, s_2, \ldots, s_n)$, let f_i denote the first STS in clone i: $f_i = \min(j : s_j \text{ hits clone } i)$. Let l_i denote the last STS in clone i: $l_i = \max(j : s_j \text{ hits clone } i)$. Let n_i denote the number of STSs hitting clone i. The cost of a permutation is the sum of its gaps: $\mathcal{C} = \sum_i (l_i - f_i - n_i + 1)$. This objective function treats gaps as false negatives. It improperly models deletions, where one event removes a series of adjacent STSs.

Karp et al. offer a more sophisticated approach to single-copy probe ordering via an approximate likelihood function. A permutation is penalized for false positives,

[1] From the same complexity results, fingerprint mapping is also hard in the worst case.

[2] Simple algebra shows this function is a metric: $d(i,j) \geq 0$; $d(i,j) = d(j,i)$; and $d(i,j) + d(j,k) \geq d(i,k)$.

false negatives, deletions, and chimeras. Simulated annealing with a modified two-opt operator is used to find likely permutations.[3] Karp et al. also investigated a simpler algorithm relying on a Hamming-distance TSP to minimize gaps. Surprisingly, its performance rivaled that of the likelihood approach in ordering simulated data. This unexpected phenomenon lead Karp et al. to conjecture that instances of the probe-ordering problem fall into one of two regimes. Instances in the first regime are characterized by sufficient information. For these, almost any reasonable algorithmic strategy will succeed in ordering probes. Instances in the second regime are characterized by insufficient information, corresponding to noisy or scarce probe-clone incidence data. No algorithm, no matter how sophisticated, can produce good probe orderings from insufficient information. If Karp's conjecture is true, the crux of probe-ordering is not algorithms, but data. [2]

The MIT-Whitehead Genome Center has used a greedy algorithm to assemble doubly-linked contigs on human chromosome 22, then ordered these contigs with an approximate likelihood function and tabu search[45]. Most recently, the advantages of using radiation hybrid mapping† for ordering single-copy probes are under investigation[18].

Arratia et al. extended the Lander and Waterman fingerprint analysis [34] to address single-copy probe mapping[4]. With clone starts and probes following independent poisson processes, their resulting model bears some resemblance to an $M/G/\infty$ queuing system [36].

5.1.2 Mapping with Multiple-Copy Probes

With single copy probes, each probe marks a single spot on the genome. Multiple copy probes do not enjoy this property.

One approach for using multiple-copy probes is to condense them into single-copy probes. The *E. Coli* project (Section 2.4) used multiple-copy restriction sites as probes, but combined them in blocks of six consecutive sites to form single-copy entities. Similarly, most fingerprinting approaches consider all of a YAC's fingerprint

bands together in an attempt to create a single-copy marker.[3]

In other situations, multiple-copy probes cannot be condensed into single-copy probes and require algorithms that explicitly address their multiplicities. Karp [2] and Newberg ([41], [40]) have developed a suite of such algorithms employing likelihood-based approaches. They model the multiple occurrences of each probe with a poisson process and assume non-chimeric clones. They employ dynamic programming to calculate the likelihood of a probe ordering in reasonable time, and they use heuristic search to find likely orderings.

The multiple-copy and single-copy probe-ordering problems are NP-complete. If one takes Karp's multiple-copy and single-copy algorithms as representative heuristics, both the multiple-copy and the single-copy probe-ordering problems appear of comparable complexity in practice. Neither problem is inherently more difficult than the other; however, they do require different algorithmic approaches. Using a single-copy algorithm to order multiple-copy probes would yield a dense tangle of spurious connections. This appears to have occurred in the CEPH-Genethon ALU-PCR mapping project.

5.2 The CEPH-Genethon ALU-PCR Map

The CEPH-Genethon ALU-PCR map is an integrated physical map of human genome combining four different types of data. The Weissenbach genetic map† [56] provides the physical map's backbone. STS† content data [15] bind genetically mapped STSs to clones. Fingerprints [8] and ALU-PCR probes ([14], [15]) establish overlap between clones. FISH data attach the genetic map to existing cytogenic maps.

These data, along with a "proposed data analysis strategy" for their use, comprise the CEPH-Genethon map [15]. The quality of this map rests on the quality of these data and the quality of this proposed strategy. Section 5.2.1 reviews the CEPH-Genethon 30 March 1994 data release, and Section 5.2.2 reviews the proposal for

[3]In contrast, Rigault discusses local mapping using individual bands as single-copy probes within a small region.[46]

their use.

5.2.1 The CEPH-Genethon Datasets

Clones

The ALU-PCR map utilizes the CEPH-Genethon megabase YAC library, first developed for the CEPH-Genethon fingerprint mapping effort (Chapters 2, 3, and 4). The library contains 33,000 clones averaging 0.91 mb in size. Over 40% of these clones are likely to be chimeric ([8], [47]); over 10% have probably deleted some portion of their DNA insert[23].

STSs

2100 markers were selected from Weissenbach's set of genetically mapped microsatellite markers to provide wide coverage of the genome. These markers were converted to STSs and screened against the YAC library using pooled testing.

FISH Data

500 YACS containing genetically mapped STSs were positioned on metaphase chromosomes using FISH, providing links between the genetic map and cytogenic maps of the human genome. The spacing between these links averaged 7.4 cM. Difficulties arising from chimerism were not discussed. It is unclear whether only clones with a single FISH localization were selected or whether only the strongest FISH localization was accepted.

Fingerprints

To a limited degree, the CEPH-Genethon `THE` and `KPN` scores discussed in Chapter 4 were used to determine clone-clone overlap. No attempt was made to remove the spikes in the fragment length distribution (Section 3.4.1). Overlaps corresponding to contaminated plates were discarded. [47]

ALU-PCR

The ALU-PCR data are the heart of the CEPH-Genethon project, for these data convert the genetic map into a physical map. This dataset consists of 6,900 ALU-PCR YAC-probes screened against a 25,000 YAC subset of the library.

ALU-PCR uses ALU-flanked primers to initiate the polymerase chain reaction†. A megabase YAC contains numerous ALUs. On average, 10 ALU pairs are close enough to sustain a PCR† reaction across the gap between them [28]. A YAC with k such pairs produces k reaction products. These k products could be separated by gel electrophoresis into k distinct ALU-PCR probes. CEPH-Genethon avoided this labor-intensive separation, choosing instead to maintain all k products in one mixed probe. Theoretically, this complex probe mixture could detect a copy of any one of the k DNA products it contains. In reality, ALU-PCR reaction products compete for amplification. Slower products may be lost due to rate kinematics. CEPH-Genethon used YAC pooling to screen each ALU-PCR probe against the library. Due to homologous regions of the genome, reaction products can occur in multiple spots across the genome. ALU-PCR probes derived from chimeric YACs also may detect multiple regions of the genome.

A simplified example presents some of the difficulties that can arise in the pooled ALU-PCR probe screenings. Figure B-31 illustrates this hypothetical case. This toy library consists of four YACs: A, B, C, and D. YAC A overlaps YAC B. YACs C and D overlap no other YACs. There are five ALU-PCR reaction products, denoted P_1, P_2, P_3, P_4, and P_5. YAC A contains $\{P_1, P_2\}$, YAC B contains $\{P_2, P_3, P_4\}$, and YAC C contains P_5 . Due to genomic homologies, YAC D also contains a copy of reaction product P_4, even though D does not overlap A or B.

YAC B is used to generate an ALU-PCR YAC probe, $Probe(B)$. Due to kinematic competition among the reaction products in PCR amplification, $Probe(B)$ contains only reaction product P_4. This probe is then screened against the library using a 2×2 pooling scheme.[4] As shown in Figure B-31, the pools are numbered 1, 2, 3,

[4]Screening four YACs with 2×2 pooling requires four tests, so this pooling saves no work. This example requires pooling, however, to demonstrate difficulties in the ALU-PCR screening

and 4 and contain YACS $\{A, C\}$, $\{B, D\}$, $\{A, B\}$, and $\{C, D\}$, respectively. Pool 1 contains reaction products $\{P_1, P_2, P_5\}$ and is not detected by *Probe*(B). Pool 2 contains $\{P_2, P_3, P_4\}$ and is detected by *Probe*(B). Pool 3 contains $\{P_1, P_2, P_3, P_4\}$ but is not detected by *Probe*(B) due to competition among the reaction products, or a false negative result. Pool 4 contains $\{P_4, P_5\}$ and is detected by *Probe*(B). From the positive hits in pools 2 and 4, it is deduced that *Probe*(B) hits YAC D. Accordingly, it is assumed that YAC B and YAC D probably overlap, and that YAC B probably does not overlap YAC A or YAC C.

This example illustrates a YAC probe that does not detect its origin clone (YAC B), that does not detect a clone it should (YAC A), and that does detect a clone it should not (YAC D). The example is somewhat contrived due to its small size. In the CEPH-Genethon ALU-PCR experiments, the pools were large (possibly 30 YACs per pool) and the YACS averaged 10 reaction products. With hundreds of reaction products in each pool, scenarios similar to this example were plausible.[28]

In the CEPH-Genethon ALU-PCR data, 40% of the YAC probes did not detect their origin clones. Competition among the ALU-PCR reactions was the likely cause. The percentage of false negative and false positive probe-clone hits cannot be obtained directly from the data; Section 5.3 approaches these rates indirectly.

Chromosomal Assignments

An ALU-PCR YAC probe derived from a chimeric YAC [7] or containing repetitive† DNA may hybridize to many locations on the genome. Recognizing that the ALU-PCR probes would not all be single-copy, CEPH-Genethon screened each probe against a panel of monochromosomal hybrids ([39], [20]) to detect multiple-copy probes. As shown in Table 5.1, the results of these screenings indicate that at least 17% of the probes are multi-copy.

A simple model gives some insight into this process. Assume each probe occurs on $X + 1$ chromosomes, where X is a Poisson random variable with mean λ. Assume that if a probe is not on a chromosome, it is never detected on that chromosome, and

---

experiments.

Number of Chromosomal Assignments	Fraction of Probes	Predicted by Simple Model, $p = 0.50$, $\lambda = 0.7$
1	47 %	48 %
2	13 %	14 %
3	3 %	2 %
4+	1 %	0 %
0 (assignment failed)	36 %	35 %

Table 5.1: Chromosomal Assignments of CEPH-Genethon ALU-PCR Probes

probe is on a chromosome, it is detected with probability p, iid.[5] As shown in the last column of Table 5.1, this model matches the observed data well for $p = 0.50$, $\lambda = 0.7$. Under these assumptions, half of the probes ($1 - e^{-0.7} = 0.5$) are multiple-copy.

5.2.2 The CEPH-Genethon Strategy

The following definitions are useful for describing the CEPH-Genethon ALU-PCR map construction strategy.

Definition 1 An ALU-PCR probe is *valid* on a path on chromosome k if

(a) the probe was uniquely assigned to chromosome k,

(b) the probe was assigned to chromosome k and other chromosomes, or

(c) the probe failed chromosomal assignment.

Definition 2 On chromosome k, two YACs *overlap* if

(a) they share an STS,

(b) at least of the YACs is an ALU-PCR probe valid for chromosome k that detects the other, or

(c) fingerprint data indicate overlap.

[5]This simple model is imperfect. The poisson model does not properly model repetitive DNA, and iid deletions do not properly model weak probes.

Definition 3 A YAC is *anchored* to a genetic locus if it is hit by an STS at the locus.

Definition 4 A *tiling path* is a minimal path of overlapping YACs between two STSs.

Definition 5 The *length* or *level* of a path is the number of YACs on the path.[6]

Definition 6 Two genetic loci are *connected at level m* if there exists a tiling path of length m or shorter between them.

Definition 7 A region of a chromosome is *covered at level m* if the region is flanked by a pair of loci which are connected at level m.

Cohen et al. proposed the following strategy to use the CEPH-Genethon ALU-PCR data to create a physical map of the human genome[15].

Rule 1 Accept the genetic map as correct.

Rule 2 Use genetically mapped STSs to anchor genetic loci to YACs.

Rule 3 Use anchored YACs and tiling paths to connect STSs within 10 cM on the same chromosome.

Rule 4 Provide a computer program, QUICKMAP[48], to provide all tiling paths of a given length between any two loci within 10 cM on the same chromosome.

Rule 5 Report coverage by chromosome and path level, as well as total genomic genetic coverage by path level.

A graph theoretic approach is helpful for understanding the CEPH-Genethon map. Consider a family of undirected graphs $G_k = (N_k, A_k)$, one for each chromosome k. A YAC is valid for chromosome k if the YAC is an ALU-PCR probe valid for chromosome k or if the YAC is not an ALU-PCR probe. For each chromosome k, the nodes of G_k are all STSs and all YACs valid for chromosome k. The arcs of G_k correspond to positive hybridizations: an arc (i, j) links STS i to YAC j if STS i hits YAC j, and an arc (l, m) links YAC l to YAC m if $Probe(l)$ hits YAC m.

[6]Only YACs are counted; for example, LOCUS 1 $\leftrightarrow$ STS 1 $\leftrightarrow$ YAC 1 $\leftrightarrow$ YAC 2 $\leftrightarrow$ YAC 3 $\leftrightarrow$ STS 2 $\leftrightarrow$ LOCUS 2 is a level 3 path.

5.2.3 Reported Results

To evaluate the CEPH-Genethon map for this thesis, the CEPH-Genethon rules were applied to the CEPH-Genethon data to replicate the CEPH-Genethon map construction process. Valid tiling paths were constructed between all pairs of STSs using breadth-first search in the G_k graphs[1]. The coverage of each chromosome and of the entire genome were computed using the CEPH-Genethon definitions. Fingerprint overlaps were excluded to focus exclusively on ALU-PCR connections. Using these rules and these data, it was found that paths of level 1, 3, 5, and 7 provided genomic coverages of 31%, 49%, 65%, and 79%, respectively.

Cohen et al. [15] reported genomic coverages of 11%, 30%, 70%, and 87% for paths of lengths of length 1, 3, 5, and 7. As Cohen's article utilized fewer probes and fewer STSs, lower coverage was expected. Lower coverage was observed for level 1 and level 3 paths. However, the coverages for level 5 and level 7 paths reported by Cohen exceed the level 5 and level 7 coverages calculated for this thesis by 5%. This anomaly might have resulted from omitting fingerprint overlaps from the map for this thesis. An alternative explanation is that CEPH-Genethon relied upon expert intervention to obtain higher genomic coverage in their map.

5.3 ALU-PCR Map Evaluation

This thesis duplicated the CEPH-Genethon ALU-PCR map creation process so that the quality of the resulting map could be investigated. Much of the resulting map appears of low quality. Section 5.3.1 presents obvious problems with the map construction strategy. Section 5.3.2 presents obvious problems with the map.

5.3.1 Problems with the CEPH-Genethon Strategy

Problem: Definition 1

The CEPH-Genethon logic behind Definition 1 is that probes which meet this criterion for chromosome k "did not contradict genetic position of the two neighboring STSs."

[15] Definition 1 is astoundingly liberal. As noted in Table 5.1, over one third of the probes lack chromosomal assignment. Definition 1 allows these probes to function as wild-card probes, valid on any chromosome. Probes with multiple assignments may be used on all chromosomes to which they were assigned. This definition indicates CEPH-Genethon's awareness that many ALU-PCR probes were multi-copy.

Problem: Definition 2b

If the ALU-PCR probes were single-copy, Definition 2b would make sense: a single shared probe would be sufficient to establish overlap between two YACs. The ALU-PCR probes, however, are not single-copy. Definition 2b applies a strategy valid for single-copy probes to multiple-copy probes; in doing so it errs gravely. As conjectured in Section 5.1.2 and demonstrated in Section 5.3.2, a dense tangle of spurious connections is the result.

Problem: Definition 4

Definition 4 is not incorrect per se: it simply inductively applies the notion of overlapping clones from Definition 2 to form paths. However, as Definition 2b creates spurious single links, Definition 4 explodes these links outward into spurious trees.

Problem: Definition 6

Definition 6 provides a very liberal definition of connected loci. Consider three adjacent ordered genetic loci, A, B, and C, with corresponding STSs a_1, a_2, b_1, c_1, c_2, c_3, and c_4.[7] Suppose a path, $\mathcal{P}$, exists between STSs a_1 and c_1, but no paths exist between any of the other 13 inter-loci STS pairs.[8] Path $\mathcal{P}$ is problematic for two reasons. First, $\mathcal{P}$ is not supported by additional paths between A and C. Second, as no path links a_1 to b_1 or b_1 to c_1, $\mathcal{P}$ passes from A to C without visiting B. Nonetheless, following Definition 6, path $\mathcal{P}$ is sufficient to cover the $A \leftrightarrow C$ genetic interval.

[7] A set of STSs defines a single locus if each STS in the set genetically maps to the same location.

[8] $\{a_1b_1, a_1c_2, a_1c_3, a_1c_4, a_2b_1, a_2c_1, a_2c_2, a_2c_3, a_2c_4, b_1c_1, b_1c_2, b_1c_3, b_1c_4\} = 13$ STS pairs.

Problem: Definition 7

As Definition 6 allows covering paths to skip STSs, Definition 7 allows covering paths to skip intervals. Consider four adjacent ordered genetic loci: A, B, C, and D. Covering the interval $A \leftrightarrow D$ is sufficient to cover the intervals $A \leftrightarrow B$, $B \leftrightarrow C$, and $C \leftrightarrow D$, regardless of their own independent coverage status.[9]

Problem: Rule 3

CEPH-Genethon provides the following rationale for Rule 3: "...by neighboring STSs, we refer to all markers located within a specific interval. Here at level 1, this interval is 10 cM.[10] This allows us to circumvent possible local inversions in the genetic map." [15] However, a 10 cM region on the Weissenbach map is considerable, representing more than local inversions of a few markers. This wide window, combined with Definitions 6 and 7, permits the coverage of large genomic regions without acknowledging skipped intervals beneath.

5.3.2 Problems with the CEPH-Genethon Map

Problem: Chromosomal Assignments

Because of Definition 1, the chromosomal assignments of the ALU-PCR probes play an important role in creating paths. Table 5.1 presented a summary of these assignments. The probe assignments may be used to provide chromosomal assignments for YACs, giving each YAC the assignments of the probes hitting it. Similarly, these derived chromosomal assignments for YACs may be used to provide chromosomal assignments for STSs, giving each STS the chromosomal assignments of the YACs it hits. In principle, each STS should possess only one chromosomal assignment. Table 5.2 presents the results of cascading chromosomal assignments up from probes through clones to STS. The last row, "ANY," denotes the full set of chromosomes.

[9] Assume A and D are within 10 cM, so Rule 3 is satisfied.

[10] Despite this sentence's insinuation to the contrary, if one accepts the computer program QUICKMAP as the definitive specification of the CEPH-Genethon ALU-PCR mapping algorithm, 10 cM is maintained as the default maximum genetic gap for paths of all lengths.

Number of Chromosomal Assignments	Observed Fraction of Probes	Implied Fraction of YACS	Implied Fraction of STSs	True Fraction of STSs
1	47 %	15 %	5 %	100 %
2	13 %	11 %	2 %	
3	3 %	9 %	1 %	
ANY	36 %	52 %	87 %	

Table 5.2: Implied Chromosomal Assignments

Such assignments are the result of "wild-card" probes, probes lacking chromosomal assignment. Table 5.2 indicates the G_k graphs are locally highly connected: 52% of all YACs and 82% of all STSs are one probe away from every chromosome. The genome is large. STSs and YACs are extremely small. STSs and YACs should not be able to reach anywhere on the genome with a path of one probe.

Problem: Bad Paths

Applying the CEPH-Genethon path construction rules to the CEPH-Genethon ALU-PCR data produces numerous bad paths. The spurious local connections in the G_k graphs explode outwards upon application of Definition 4, becoming tangled forests of spurious trees. Figure B-32 illustrates this for one probe, `706d7`. Probe `706d7` can reach fourteen chromosomes in paths of 3 YACs or less. Note that every path conforms to the CEPH-Genethon rules. Figure B-33 presents a histogram of the number of chromosomes each ALU-PCR probe can reach with valid paths of three YACs or less. For example, probe `706d7` is one of the probes represented by the histogram bar at 11. Figure B-33 indicates that probe `706d7` is not exceptional in its ability to reach many chromosomes in three YACs; many probes are within three YACs of many chromosomes.

If many probes are close to many chromosomes, it is likely many probes are near many other probes. To investigate this, all same-chromosome STS pairs were grouped by inter-STS genetic distance. The cumulative fractions of connected pairs as a function of path length for each group were calculated. These are presented in

Figure B-34. This potent diagram raises four issues.

First, note the lowest curve corresponds to STS pairs greater than 50 cM apart. A generally accepted rough equivalence is one centiMorgan equals one megabase.[44] Though recombinational hot-spots or unusually long YACs may change this ratio slightly, this equivalence corresponds to a ratio of one YAC for each centiMorgan.[11] A path spanning 50 cM using only ten YACs is almost certainly spurious. As shown in Figure B-34, such paths abound in the G_k. Indeed, almost 80% of the 46000 STS pairs greater than 50 cM apart are bridged by paths of 10 or fewer YACs.

Second, for longer paths, the number of paths appears independent of genetic distance. The curves corresponding to STS pairs within 5–10 cM, 10–20 cM, and 20–50 cM fall nearly on top of the 50+ cM curve. If all paths were valid, one would anticipate that short paths would connect a greater fraction of shorter genetic gaps than longer genetic gaps. This is not case for paths above 5 cM, indicating they are largely spurious. The anticipated pattern is observed for shorter gaps, suggesting than the set of paths linking STSs within 5 cM contains a significant fraction of valid paths.

Third, if the lowest curve of Figure B-34 corresponds to spurious paths, and if the spurious path process is independent of inter-STS distance, then this curve provides an estimate of the spurious path rate. In turn, this rate may be used to calculate the expected number of incorrect paths for each length and distance. By subtracting the expected number of incorrect paths from the observed number of paths, Figure B-34 may be adjusted to present the expected cumulative fraction of valid paths.[12] Figure B-35 plots the results of these calculations. Again, very few paths connecting STSs more than 5 cM apart are valid. Further, for STSs within 5 cM, most true paths involve four or fewer YACs. Accepting longer paths, even between close STSs, does not bring in many more valid paths.

---

[11]This conversion is conservative for two reasons. First, YACs average less than 1 mb in length. Second, adjacent YACs on paths must overlap, making the genetic distance they cover smaller than the sum of their lengths.

[12]If the expected number of incorrect paths exceeded the number of paths observed, all of the observed paths were designated incorrect: $E(n_{valid}) \equiv \max(0, n_{total} - E(n_{invalid}))$. From this definition, the 50+ cM curve in Figure B-35 is identically zero.

Fourth, Figure B-34 indicates the number of spurious paths (most paths linking STSs more than 5 cM apart) explodes rapidly after length three or four. This rapid phase-transition from low to high connectivity is one characteristic of random graphs [11]. The G_k are highly structured, not random, graphs. Nonetheless, the G_k do possess spurious ALU-PCR connections. The superposition of these bad arcs onto the "true" G_k might yield graphs that behave like their random cousins.

Problem: Random Graph Behavior

A recent award-winning play, "Six Degrees of Separation," was named for the idea that any two people know one another through a chain of six or fewer acquaintances. Random graphs are known to have an analogous property: relatively short paths connect most pairs of nodes.[10] Figures B-34 and B-35 indicate the G_k may have this property as well.

To investigate the difference between the behavior of the G_k and the behavior of random graphs, another set of graphs, the $\tilde{G}_k$, was assembled in the following way. The CEPH-Genethon STS content data, the CEPH-Genethon genetic map, and the CEPH-Genethon chromosomal assignments of ALU-PCR probes were left unchanged. The ALU-PCR data, however, were replaced with random data, preserving the correct number of hits per ALU-PCR probe. For example, in the true data, ALU-PCR probe `100g7` was assigned to chromosomes 9, 15, 21, and 22, and hit fourteen YACs: `734b10`, `745g7`, `770d6`, `784e9`, `800a4`, `801a4`, `829d2`, `829d8`, `830e12`, `878c5`, `899g9`, `902g8`, `921d9`, and `928e3`. In the scrambled data, ALU-PCR probe `100g7` remained assigned to chromosomes 9, 15, 21, and 22, and still hit fourteen YACs. Now, however, the fourteen YACs represent fourteen random draws from all the clones in the library.[13] These fourteen random YACs were `969e12`, `639e8`, `882e2`, `983f11`, `968d9`, `899d10`, `919c7`, `907b4`, `865d9`, `689h5`, `756c6`, `684h9`, `779g4`, and `926e2`.

Figure B-36 replicates Figure B-34 for the $\tilde{G}_k$, plotting the fraction of connected STS pairs as a function of distance and path length. Figure B-36 is strikingly similar

[13]The YACs were independently drawn from the library of 25000 YACs, with replacement, using a random number generator.

to Figure B-34. The primary difference between the two plots is that the G_k contain more short paths between near STSs than the $\tilde{G}_k$. Aside from paths between STSs within 5 cM involving fewer than four YACs, with respect to connections between STS pairs, the G_k behave like the $\tilde{G}_k$. In short, for most paths, the CEPH-Genethon ALU-PCR map behaves like a random graph. This provides further evidence that most of the paths longer than 5 cM and involving more than four YACs are invalid.

Figure B-37 presents genomic coverage using the CEPH-Genethonmap construction rules for G_k and $\tilde{G}_k$. The first 30% of this coverage is known to be correct, for paths of length one correspond to STS content alone. Coverage in the $\tilde{G}_k$ increases slowly for paths of length two and three, then rapidly increases at path length four. By length eight, the scrambled data cover as much of the genome as the real data. In contrast, the coverage in real G_k increases more substantially for paths of lengths two and three. This differential corresponds to the valid, short paths present in the real data but missing in the scrambled data. Subsequent increases correspond to bad paths present in both the real and the scrambled data.

Summary

The CEPH-Genethon map construction rules applied to the CEPH-Genethon ALU-PCR data produce a map of poor quality. Definitions 1 and 2b treat multiple-copy ALU-PCR probes as single-copy entities, creating many false connections. Definition 4 blows up these spurious paths into spurious trees. Rule 3 and Definitions 6 and 7 allow paths to skip intermediate markers and intervals. Most valid paths involve fewer than four YACs and span less than 5 cM. With respect to connectivity and coverage, the CEPH-Genethon map resembles a random graph.

5.4 ALU-PCR Map Remedies

The CEPH-Genethon ALU-PCR data are an invaluable resource for the human genome research community. Nonetheless, the strategy that accompanies these data leaves much to be desired.

CEPH-Genethon realized the shortcomings of this early map. They have opted to use it only as scaffolding for further STS screenings.[14] With a two- or three-fold increase in the number of STSs, most intervals on the map would be covered by a level one path. This would yield a pure STS-content map, a physical map that did not rely on questionable ALU-PCR connections.

If used in a more prudent fashion, the ALU-PCR data can still prove useful to the community before additional STSs are added. This section explores three conservative strategies for using these data. The first alternative, `Within5cM`, modifies only Rule 3. The second alternative, `NoWildcardProbes`, modifies only Definition 1. The third, `Win5/NoWild`, modifies both Rule 3 and Definition 1.

5.4.1 Alternative Strategy: `Within5cM`

Rule 3 uses a 10 cM window to define close genetic loci. As shown in Figure B-34, this window is too liberal. The `Within5cM` strategy adopts all of the CEPH-Genethon rules but modifies Rule 3.

Rule 3' Use anchored YACs and tiling paths to connect genetic loci within 5 cM on the same chromosome.

Table 5.3 presents the genomic coverage that results from the `Within5cM` strategy and the coverage that results from the original `Within10cM` strategy. Figure B-38 plots these two measures and their difference. For paths of length one and two, `Within5cM` enjoys almost the same coverage as `Within10cM`. For paths of length three and longer, this difference increases, indicating that for longer paths the more restrictive `Within5cM` strategy reduces genomic coverage considerably.

Using the same scrambled data as before, the $\tilde{G}_k$ graphs were analyzed under the `Within5cM` strategy. Figure B-39, analogous to Figure B-37 for the original `Within10cM` strategy, plots real and scrambled coverage against path length for the

[14]They are not anxious to perform this work themselves, however, and have asked the international community—in particular, the MIT Genome Center—for assistance.[47]

Path Length	Original Strategy: `Within10cM`	Alternative Strategy: `Within5cM`
1	31%	29%
2	39%	36%
3	49%	42%
4	57%	46%
6	65%	49%
7	74%	54%
8	80%	62%

Table 5.3: Genomic Coverage

map generated with `Within5cM`. If the `Within5cM` strategy served to eliminates spurious linking paths completely, the curve corresponding to scrambled coverage would be essentially flat. As shown in Figure B-39, the observed curve is not flat. As `Within5cM` permits spurious paths in the scrambled graphs, it most likely permits spurious connections in the real graphs as well. This suggests the G_k will also contain spurious paths at longer lengths under the `Within5cM` strategy. Coverage in these scrambled graphs, however, increases more slowly with path length than in the original `Within10cM` G_k graphs. This suggests that `Within5cM` removed some, but not all, of the problems in the CEPH-Genethon map.

Adopting the `Within5cM` strategy and using only paths of three YACs or shorter produces a total genomic coverage of 42%. This is less than half the of the 87% coverage reported by Cohen et al. [15]

5.4.2 Alternative Strategy: `NoWildcardProbes`

Table 5.1 indicates that 36% of the ALU-PCR probes failed chromosomal assignment. Under Definition 1c, such probes may be used to form paths on any chromosome. To investigate the effect of these wild-card probes, the `NoWildcardProbes` strategy prohibits their use entirely. All of the other CEPH-Genethon path construction rules were adopted without change. In particular, the `Within10cM` strategy was used.

Definition 1" An ALU-PCR probe is *valid* on a path on chromosome k if

Path Length	Original Strategy: `UseWildcardProbes`	Alternative Strategy: `NoWildcardProbes`
1	31%	27%
2	39%	34%
3	49%	41%
4	57%	44%
6	65%	48%
7	74%	50%
8	80%	56%

Table 5.4: Genomic Coverage

(a) the probe was uniquely assigned to chromosome k, or

(b) the probe was assigned to chromosome k and other chromosomes.

(c') Probes failing chromosomal assignment are not valid for any chromosome.

Note the `NoWildcardProbes`strategy discards 36% of the ALU-PCR data. Table 5.4 compares the genomic coverage that resulted from the `NoWildcardProbes` strategy with the coverage from the original strategy, `UseWildcardProbes`. Figure B-40 plots these same coverage statistics. As expected, the more restrictive `NoWildcardProbes` strategy produces lower genomic coverage. However, the difference between `UseWildcardProbes` and `NoWildcardProbes` remains small until paths of length two or three are reached. This again supports the earlier conclusion that longer paths are of suspect quality. It also supports the notion that `NoWildcardProbes` was able to remove some bad paths without removing good ones (the shorter ones).

Figure B-41 plots the cumulative fraction of connected STS pairs as a function of path length for each distance group for the `NoWildcardProbes` strategy. This figure is quite different than Figure B-34, the analogous plot for the original `UseWildcardProbes` approach. The highest curve in Figure B-41 corresponds to intra-locus STS pairs. For the `NoWildcardProbes` strategy, the curve rises quickly from 38% at level one to 53% at level four. It then rises slowly to 60% at level ten and levels out. This indicates that the `NoWildcardProbes` strategy has curtailed the

rapid explosion of intra-loci connections between level six and level nine that occurred with the `UseWildcardProbes` strategy.

Figure B-43 plots the genomic coverage obtained using the `NoWildcardProbes` strategy on the scrambled $\tilde{G}_k$ data and the original G_k real data. Under the original `UseWildcardProbes` strategy, the scrambled data enjoyed the same coverage as the real data by level seven, as shown in Figure B-37. In contrast, under `NoWildcardProbes`, the coverage of the scrambled data remains below the coverage of the real data.[15] The lower curve in Figure B-43 corresponds to coverage in the $\tilde{G}_k$, or spurious coverage. This curve increases with path length, though it increases slowly. This indicates the `NoWildcardProbes` strategy was unable to curtail all invalid paths.

An effective strategy would allow the coverage to increase with path length in the real data without a corresponding increase in the scrambled data. Such a situation would provide some confidence that the new paths gained with increasing path length were valid.

5.4.3 Alternative Strategy: `Win5/NoWild`

The third alternative strategy, `Win5/NoWild`, combines the both of previous approaches to achieve this goal.

Definition 1"' An ALU-PCR probe is *valid* on a path on chromosome k if

(a) the probe was uniquely assigned to chromosome k, or

(b) the probe was assigned to chromosome k and other chromosomes.

(c"') Probes failing chromosomal assignment are not valid for any chromosome.

Rule 3"' Use anchored YACs and tiling paths to connect genetic loci within 5 cM on the same chromosome.

---

[15] Precisely, the scrambled data yield lower coverage than the real data for paths of length eight or less. It is possible these two curves meet at some length greater than eight.

Figure B-44 presents coverage for this strategy. The lower curves corresponds to the scrambled data. This curve remains effectively flat, rising only 5% from 29% at level one to 35% by level eight. The upper curve corresponds to paths in the real data. This curve rises more quickly, reaching 42% coverage by level eight. The difference in coverage reflects the valid paths that exist in the G_k that do not exist in the scrambled $\tilde{G}_k$.

Analogous to Figures B-34 and B-35, Figure B-42 adjusts Figure B-41 by subtracting off the expected number of false paths for the `NoWildcardProbes` strategy. The lowest curve, pairs 50+ cM apart, is zero by definition. Assuming most short paths between STSs 20–50 cM pairs apart were also spurious, Figure B-42 indicates that the G_k remain relatively free of spurious paths until level seven. For the sets of STSs 0 cM, 0–2 cM, and 2–5 cM apart, the number of short connecting paths grows until level four. After level four, the upper three curves in Figure B-42 level off.

Using the `Win5/NoWild` strategy, the G_k remain relatively free of likely-to-be-spurious paths until level seven. The number of likely-to-be-correct short connecting paths increases until level four, then levels off. This is an ideal situation: using `Win5/NoWild` and paths no longer than four appears to have provided most of the correct connections while avoiding most of the incorrect ones. This combined strategy covered 38% of the genome reliably, considerably lower than CEPH-Genethon's original reported 87% coverage.

The `Within5cM` strategy and level one paths covered 29% of the genome. Level one paths correspond to STS-content mapping. After imposing the stringent `Win5/NoWild` filter on the CEPH-Genethon ALU-PCR map to avoid unreliable paths, only 38% of the genome is covered. The costly ALU-PCR data provide only an additional 9% of reliable coverage. This analysis suggests that screening additional STSs instead of ALU-PCR probes might have provided CEPH-Genethon with superior coverage at a lower cost.

5.5 Conclusion

This chapter has examined one recent probe-based mapping effort, the CEPH-Genethon ALU-PCR project. This project used ALU-PCR probes to establish overlap between YACs in the CEPH-Genethon megabase library. The project used these overlapping YACs to bridge the intervals between genetically mapped STSs.

The ALU-PCR probes were not single-copy, as shown by the chromosomal assignment data. Nonetheless, CEPH-Genethon adopted a map construction strategy that assumed single-copy probes. As a result, the CEPH-Genethon map contained many spurious connections and paths. Both close and distant loci were similarly connected by short paths. In terms of coverage and connectivity, the ALU-PCR graph had similar behavior to a random graph.

The strategy `Win5/NoWild` prohibits the use of wild-card probes and only allows paths between loci within 5 cM. This approach appears to yield reliable paths up to length four. When applied to the CEPH-Genethon data, this approach produced a map covering 38% of the human genome. Most of this coverage was the result of level one paths, where the ALU-PCR data played no role. In contrast, the original strategy proposed by CEPH-Genethon extensively used ALU-PCR wild-card probes to cover 87% of the genome unreliably.

Chapter 6

Conclusion

This thesis has examined two methods for constructing genomic physical maps: fingerprint mapping and hybridization probe mapping.

Fingerprinting methods enjoyed considerable success on projects of moderate size. The *E. Coli* mapping effort is one such example. The method appears less suitable for larger genomes. CEPH-Genethon undertook an ambitious set of experiments to map the entire human genome with fingerprints. This fingerprinting effort cannot be called a success. The project did not achieve its goal: it did not map the human genome. In addition, these costly fingerprint data were library-specific. Unlike STS-content mapping, which is based permanent sequence-based landmarks, the fingerprint pattern data apply only to the CEPH-Genethon megabase YACs. This library is chimeric and unstable. Better cloning systems, such as BACs, are preferable for sequencing, the next stage of the genome project. For the these reasons, it is likely the CEPH-Genethon megabase YAC library will be discarded within the next five years. At this point, the CEPH-Genethon fingerprint data will have no value. Even today, the data appear riddled with errors. It is probable additional data errors remain undetected. Nonetheless, the CEPH-Genethon `THE` and `KPN` overlap statistics performed admirably for detecting pairwise overlap among the CEPH-Genethon YACs, given the poor finerprint data quality.

This thesis represents the first independent use of the CEPH-Genethon fingerprint data and the first independent assessment of the `THE` and `KPN` statistics. Three new

statistics were proposed and implemented. The best of these, the `Entropy` statistic, was slighlty less powerful than `THE`. However, even the simplest statistic, `Trinomial`, performed within an order of magnitude of the elaborate CEPH-Genethon tests. This suggests the limiting feature for detecting YAC overlap using the CEPH-Genethon fingerprints is not algorithms, but data quality.

The `Entropy` statistic has proved useful in STS-YAC address disambiguation. This use of the statistic has cut the number of disambiguation re-tests at the MIT-Whitehead Genome in half, saving time and resources.

The last chapter of this thesis presented a first independent assessment of the much-heralded CEPH-Genethon ALU-PCR map. This analysis indicates that the CEPH-Genethon strategy for map construction is far too liberal. When applied to the CEPH-Genethon ALU-PCR data, almost all of the the linking paths that result are spurious.

This thesis provided a more conservative strategy, `Win5/NoWild`, for the CEPH-Genethon data. This approach produced paths that appear mostly valid. On the other hand, using the data in this manner drops genomic coverage from the 87% reported by CEPH-Genethon to only 38%. The first 29% of this coverage was obtained solely through STS-content and did not involve ALU-PCR overlaps at all.

Generating the fingerprint and ALU-PCR data must have been a costly and laborious process for CEPH-Genethon. This thesis suggests neither dataset justified these expenditures. While a thesis composed of negative assessments of another's data is mildly disheartening, it this process of independent verification that ensures the honesty and integrity of scientific endeavors.

Appendix A

Glossary

† **A**

ALU A repetitive sequence of DNA ubiquitous on the human genome.

ALU PCR A polymerase chain reaction using ALU-flanked primers.

ambiguous STS hit Due to pooled testing, one or more hybridizations may indicate a set of clones, of which one or more contains the STS. The STS is said to "hit ambiguously" each candidate clone in the set.

† **B**

basepair A single nucleotide. The quantum unit of distance in physical mapping. Abbreviated "bp".

† **C**

CEPH The Centre d'Etude du Polymorphisme Humain. A research institution in Paris.

chimera A clone with multiple DNA inserts spliced together.

chromosome A complete DNA molecule, confined to the cellular nucleus, containing coding and non-coding regions.

clone A replicating entity containing inserted DNA of interest.

complete restriction digest A restriction digest in which the DNA strand is cut at every recognition site.

clone library A collection of clones.

contig Two or more overlapping clones that cover a contiguous region of the genome. May refer to an ordered or unordered collection of clones, depending on context.

cytogenic location A location on a metaphase chromosome, typically denoted by a band position and an arm designation (p or q).

† **D**

deletion The process of a YAC losing some or all of its insert. Certain regions of the genome are thought to be more likely to be deleted when cloned into YACS.

diploid chromosome set Both members of each set of chromosome pairs. The normal human diploid chromosome set contains 46 chromosomes (44 autosomes and two sex chromosomes).

double linkage Two clones are "doubly linked" if they share two STSs. Double linkage is one technique to avoid false links between STSs caused by chimeric clones.

double restriction digest A restriction digest using two restriction enzymes at once.

† **E**

electrophoresis See "gel electrophoresis."

enzyme A biological catalyst, often a protein.

† F

FISH Fluorescent In-Situ Hybridization. A technique to visualize a probe hybridizing to a metaphase chromosome. Allows probe mapping at the cytogenic band level of resolution.

four-cutter A restriction enzyme with a 4 bp recognition site.

† G

gel electrophoresis A technique to size tagged fragments of DNA using an electric field (fixed or variable) and a gel medium. The electric force pulls fragments though the medium, where a fragment's velocity is a decreasing function of its length. After a fixed time, one uses the tag (often radioactivity or chemiluminescence) to determine the position of each fragment within the gel; these final positions indicate fragment size.

genetic algorithm A randomized heuristic optimization technique for combinatorial problems. Introduced by Holland in [27].

genetic map A map in which markers are linearly positioned along a chromosome by natural crossover events in meiosis.

genome The complete DNA of an organism.

$(GT)_n$ probe A probe that detects the repetitive sequence "GTGT ... GT".

† H

haploid chromosome set One (randomly selected) member of each set of chromosome pairs. The normal human haploid chromosome set contains 23 chromosomes (22 autosomes and a sex chromosome).

hybridize The process hydrogen bonding of two single-stranded segments of DNA, or of one single-stranded segment of DNA and one segment of RNA with comple-

mentary sequences. A mechanism for discriminating, among many fragments, those containing a segment of interest.

† **I**

inter-ALU PCR A polymerase chain reaction using ALU-flanked primers.

island Singleton clones and contigs covering the genome. Introduced in [34].

† **K**

kilobase 10^3 basepairs. Abbreviated "kb".

† **M**

megabase 10^6 basepairs. Abbreviated "mb".

metaphase chromosome A stage of cellular replication in which the chromosomes condense into characteristic "X" arrangements, visible with light microscopy.

multiple restriction digest A restriction digest using several restriction enzymes at once.

† **N**

nucleotide A single base. The quantum unit of distance in physical mapping.

† **O**

oceans Genomic regions covered by no clones. Introduced in [34].

† **P**

PCR Polymerase Chain Reaction. An efficient mechanism for rapid amplification of DNA between two primer sequences.

physical map A linear positioning of markers along a chromosome in terms of physical distances (for example, kilobases).

probe An assayable marker on the genome.

† **R**

radiation hybrid mapping A physical mapping method analogous to genetic mapping in which random breaks in chromosomes subjected to radiation play the role of crossovers in meiosis.

random-base DNA model A crude model of DNA sequence, in which the four bases each occur with probability $\frac{1}{4}$, independent and identically distributed, at every site along the genome.

recognition site The particular sequence at which a given restriction enzyme cuts.

repetitive sequence A DNA sequence that repeats many times across a genome. ALU is one such example on the human genome.

restriction digest The resulting inter-recognition site fragments obtained by cutting a strand of DNA with a restriction enzyme.

restriction enzyme An enzyme that cuts DNA at a specific recognition site.

restriction map The spacing of restriction sites along a region of DNA.

† **S**

single linkage Two clones are "singly linked" if they have both have definite STS hit in common.

six-cutter A restriction enzyme with a 6 bp recognition site.

STS Sequence Tagged Site. Typically, short sequences of DNA (100–500 bp) assayable by flanking PCR primers. Usually selected to be single copy. Proposed in [42].

single restriction digest A restriction digest using a single restriction enzyme.

† **U**

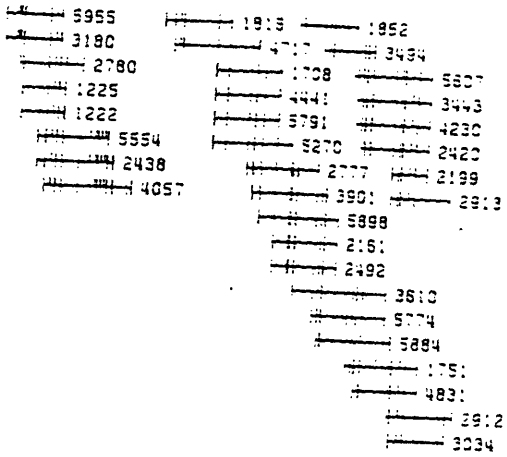
unambiguous STS hit One or more hybridizations that indicate an unique clone positive for the STS. (cf. "ambiguous STS hit")

† **Y**

YAC Yeast Artificial Chromosome, a type of clone. YACS permit large inserts (500 kb - 2000 kb) but are often chimeric and delete.

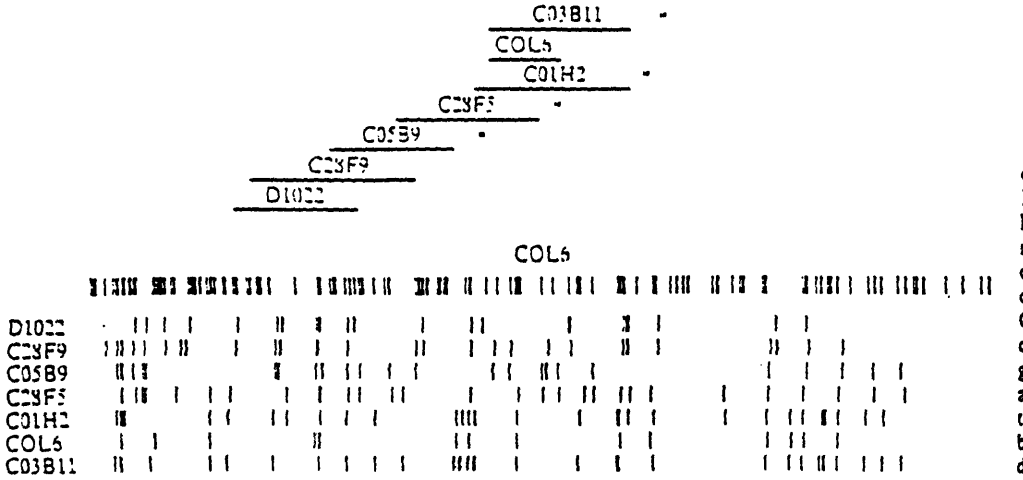
Appendix B

Figures



A representative map unit spanning the *CYC7–CEN5* region on chromosome V. In both the composite map and the individual-clone maps, vertical lines that extend both above and below the horizontal line represent group boundaries, while lines that extend only above the horizontal line demarcate the RH fragments within a group. RH fragments within a group are arbitrarily mapped in order of decreasing size from left to right.

Figure B-1: Yeast Map Sample (Olson et al. 1986)



A contig displayed on the computer screen. (*Upper*) Each clone is represented by a line of length proportional to the number of bands. Asterisk indicates the presence of hidden clones. Repeat of the name COL6 beneath the contig indicates location of this known gene; additional remarks can be added as required. (*Lower*) Pattern of marker bands and clone bands, plotted from digitized data.

Figure B-2: Worm Map Sample (Coulson et al. 1986)

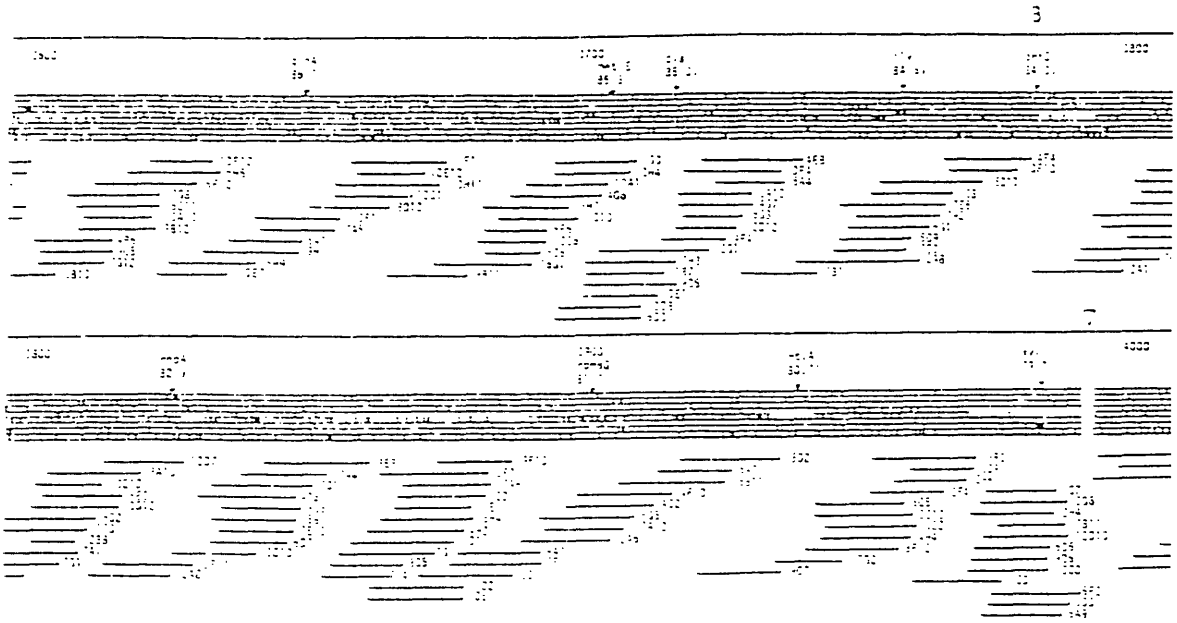
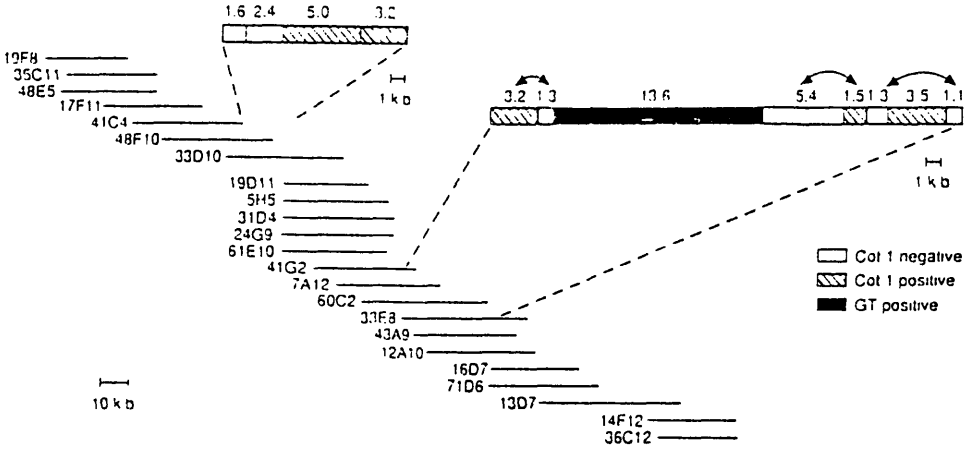


Figure B-3: Bacterium Map Sample (Kohara et al. 1987)



. An example of a 240-kb contig (no. 33) obtained after the analysis of GT-positive cosmid clones. The expanded maps at the top show that some ordering of restriction fragments (*Eco*RI in this case) can be obtained based on differences in overlap. Information as to whether the fragments contain $(GT)_n$ sequences, other repeat sequences, or single-copy DNA is also obtained.

Figure B-4: Chromosome 16 Map Sample (Stallings et al. 1990)

	Yeast 1986	Worm 1986	Bacterium 1987	Chromi6 1990	Human 1992
Digestion Approach	[illegible] double 6-cutter, complete	6-cutter then 4-cutter, complete	[illegible] single 6-cutters, partial	[illegible] single 6-cutters, [illegible] double 6-cutter, complete	[illegible] single 6-cutters, complete
Cloning Vector	Cosmid	Cosmid and Lambda	Lambda	Lambda	YACS
Chimerism	none reported	none reported	none reported	some	[illegible]
Bands Measured	all	all	all	all	bands with probes
Probes	none	none	none	two for each band	bands with probes
Data Entry	manual	semi-manual	manual	semi-manual	semi-automated
Bayesian Approach	no	no	no	yes	yes
Band Matching	most likely	most likely	not applicable	all matchings	all matchings
Overlap Statistic	heuristic	binomial	shotgun sequence assembly	elaborate model	elaborate model
Reported Coverage	95%	[illegible]	[illegible]	[illegible]	[illegible]
Qualitative Assessment of Contigs	numerous and small	numerous and small	few and large	numerous and small	numerous and very small
Published Map	yes	yes	yes	partial	no

Figure B-5: Mapping Project Features

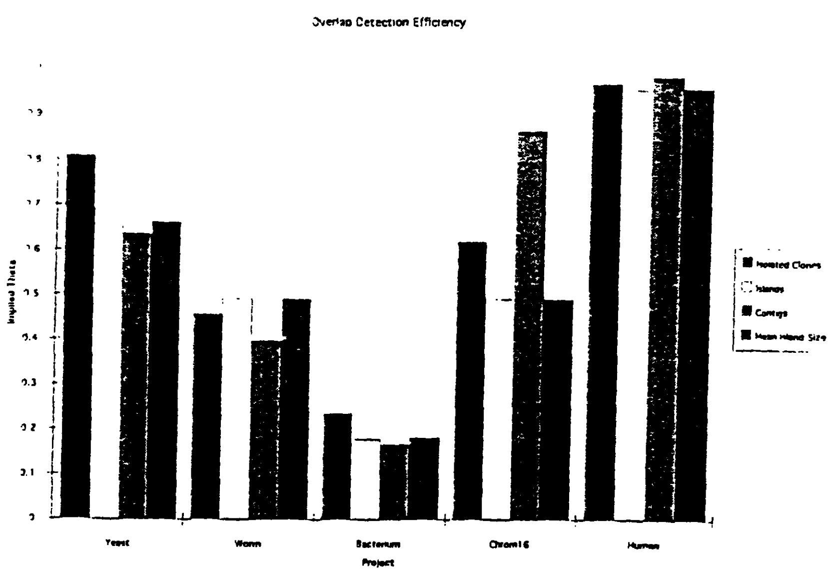


Figure B-6: Implied Thetas, Five Projects

```
Clone 889_a_1 :
enzyme EcoRI : Bands 4
 9217 0.887 8453 0.818 7385 0.372 3491 0.151
enzyme PstI : Bands 4
 12565 0.443 7653 0.289 6653 0.108 5847 0.112
enzyme PvuII : Bands 5
 11503 0.334 7039 0.422 5770 0.526 3947 0.114 3160 0.114
Clone 889_a_2 :
enzyme EcoRI : Bands 1
 17680 0.505
enzyme PstI : Bands 1
 5687 0.394
enzyme PvuII : Bands 1
 2622 0.917
```

Figure B-7: Original CEPH Data Format

```
Clone 889_a_1 :
 probe Kpn :
enzyme EcoRI : Bands 4
 9217 8453 7385 3491
enzyme PstI : Bands 4
 12565 7653 6653 5847
enzyme PvuII : Bands 5
 11503 7039 5770 3947 3160
 probe THE :
enzyme EcoRI : Bands 10
 16291 11769 10372 8892 6935 5901 5651 2554 2240 2070
enzyme PstI : Bands 7
 19295 12661 11117 2548 1992 1890 1591
enzyme PvuII : Bands 8
 25076 17009 10738 8031 5159 4194 3626 2346
Clone 889_a_2 :
 probe Kpn :
enzyme EcoRI : Bands 1
 17680
enzyme PstI : Bands 1
 5687
enzyme PvuII : Bands 1
 2622
 probe THE :
enzyme EcoRI : Bands 0

enzyme PstI : Bands 0

enzyme PvuII : Bands 0
```

Figure B-8: Current CEPH Data Format

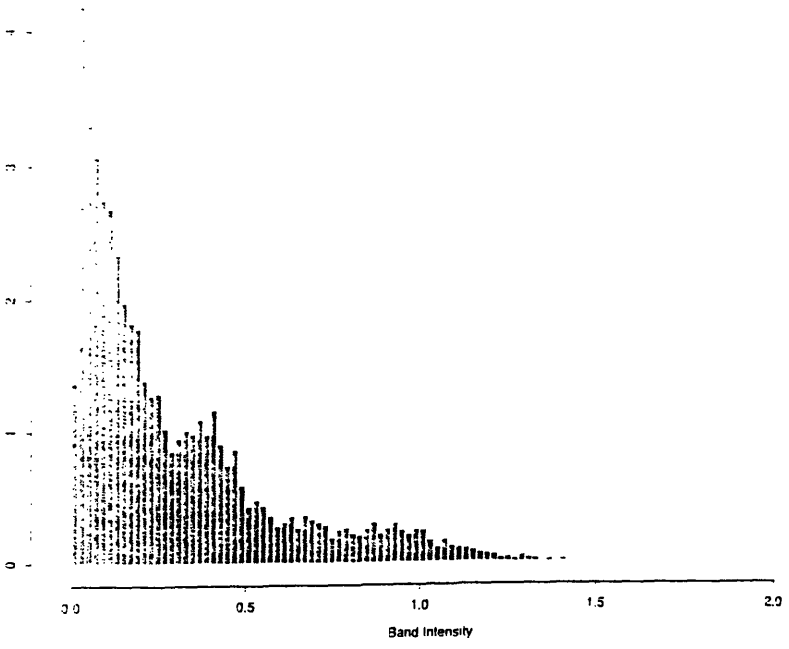


Figure B-9: Optical Intensities of CEPH Fingerprint Bands

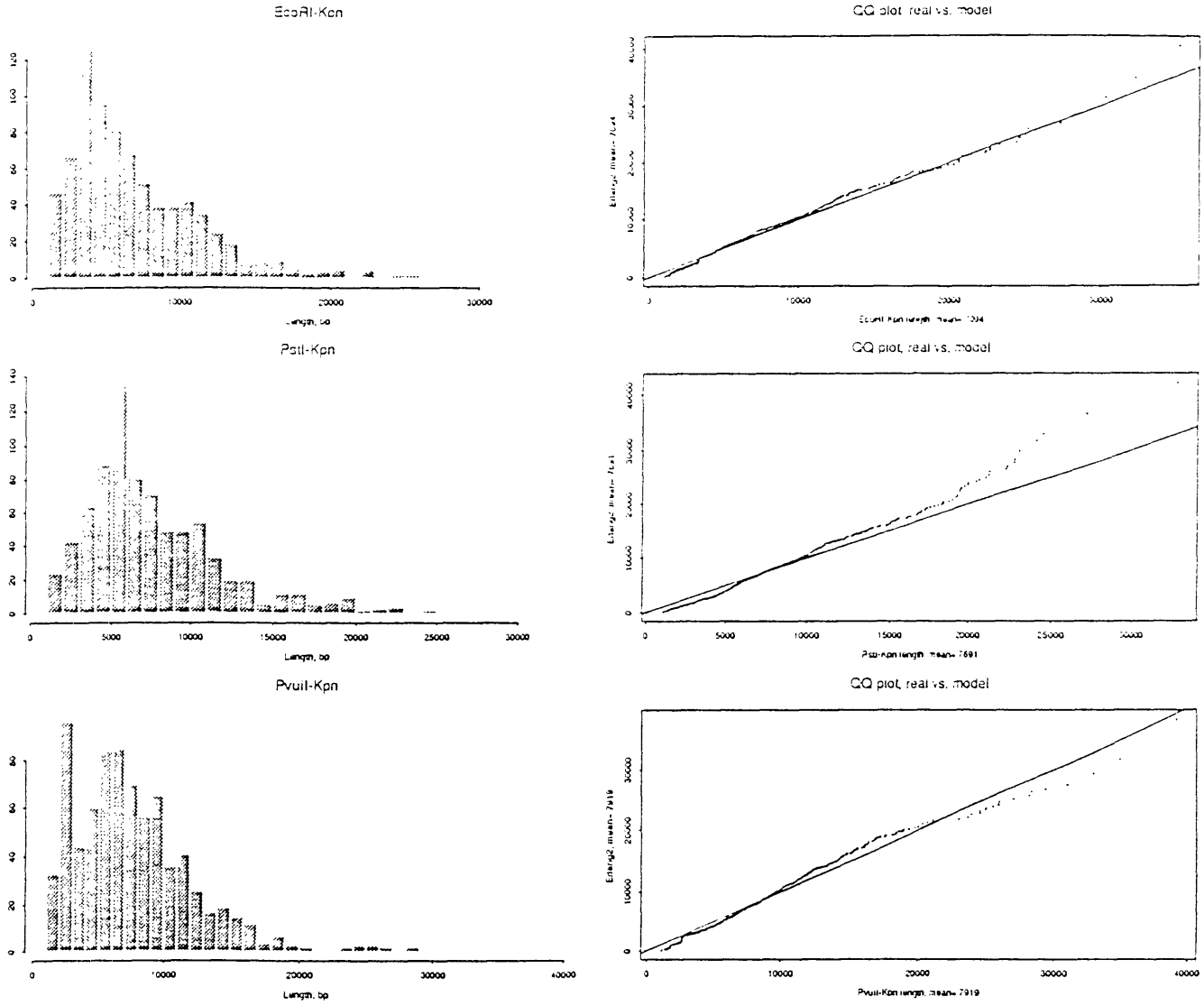


Figure B-10: CEPH Fingerprint Band Size, KPN

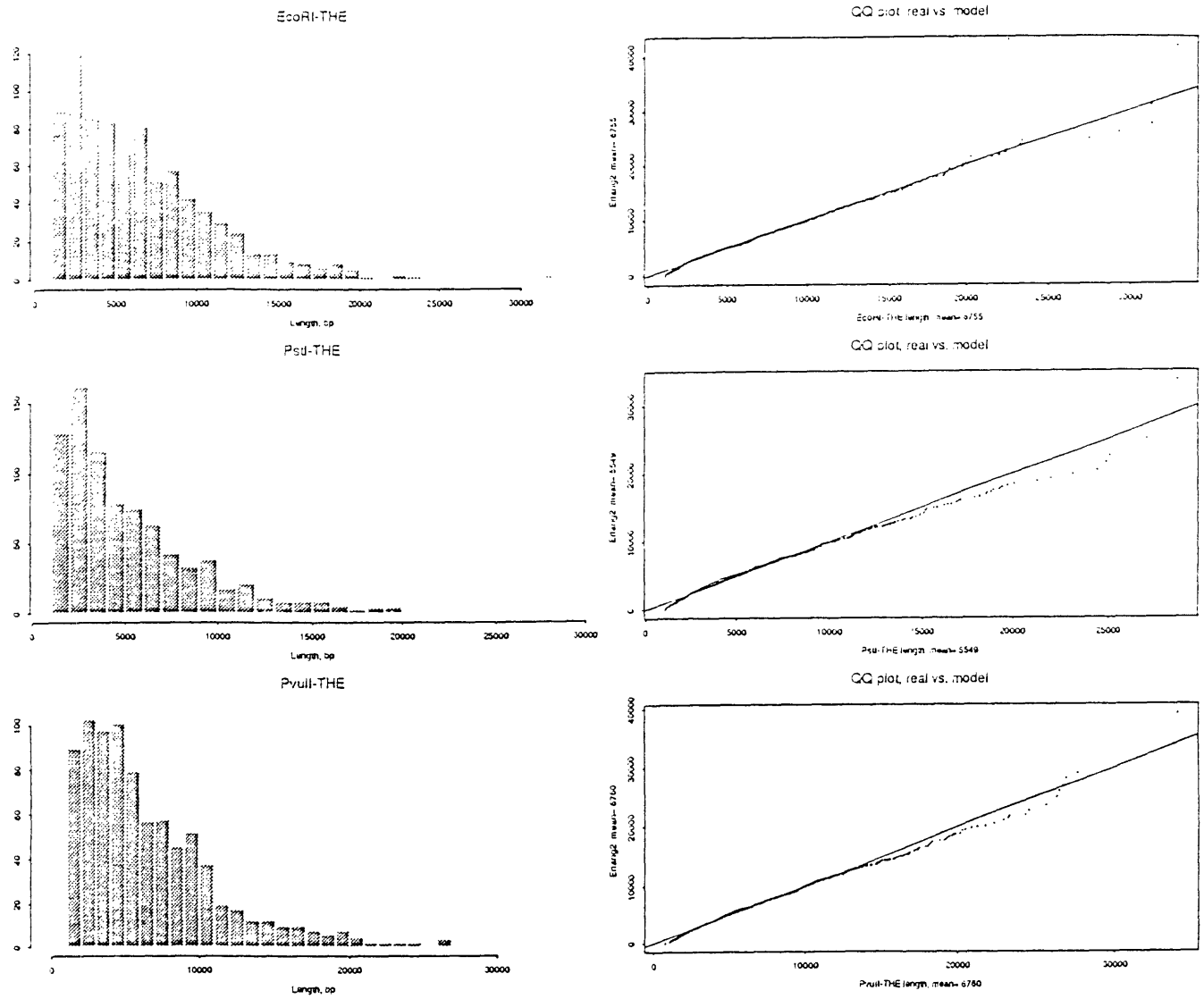


Figure B-11: CEPH Fingerprint Band Size, THE

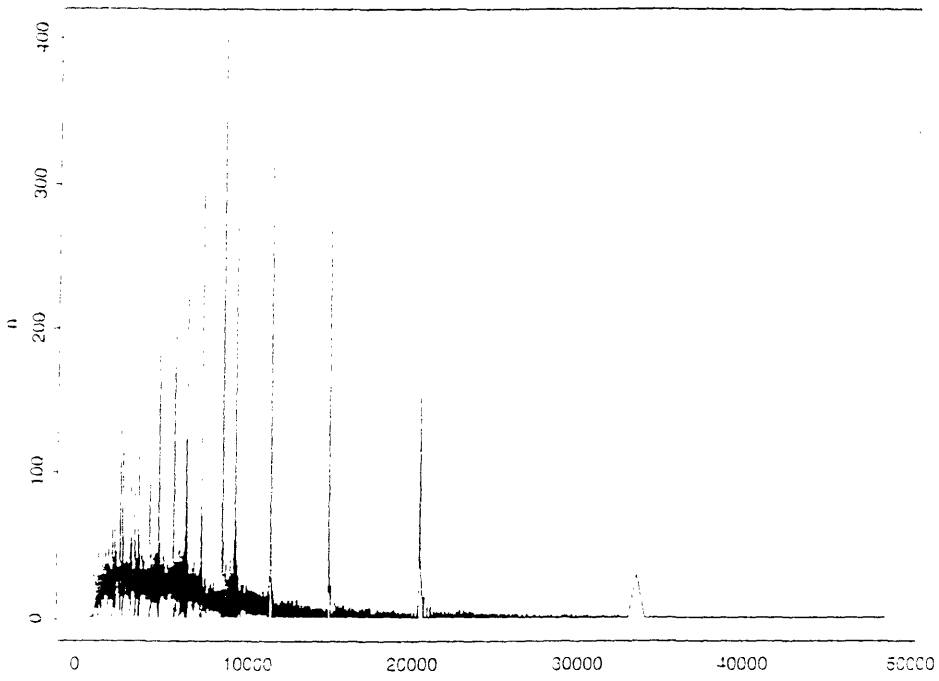


Figure B-12: Fine Histogram of CEPH Fingerprint Band Size

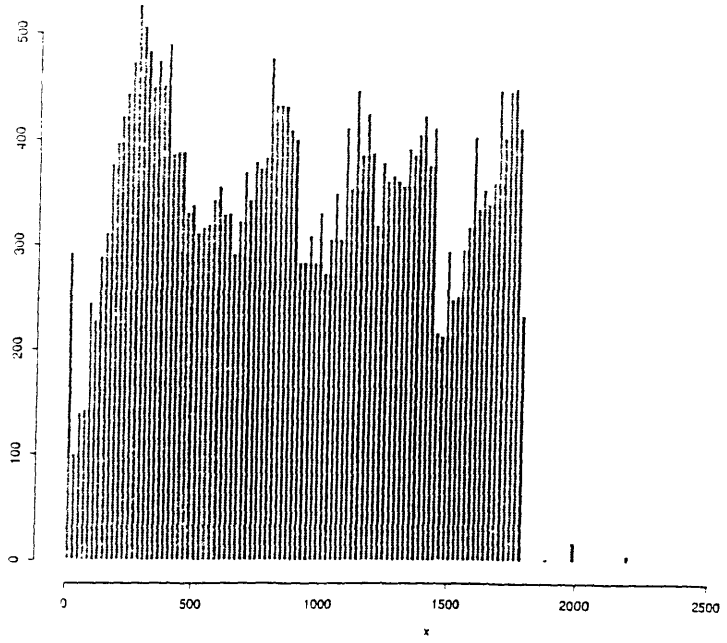


Figure B-13: YAC Lengths

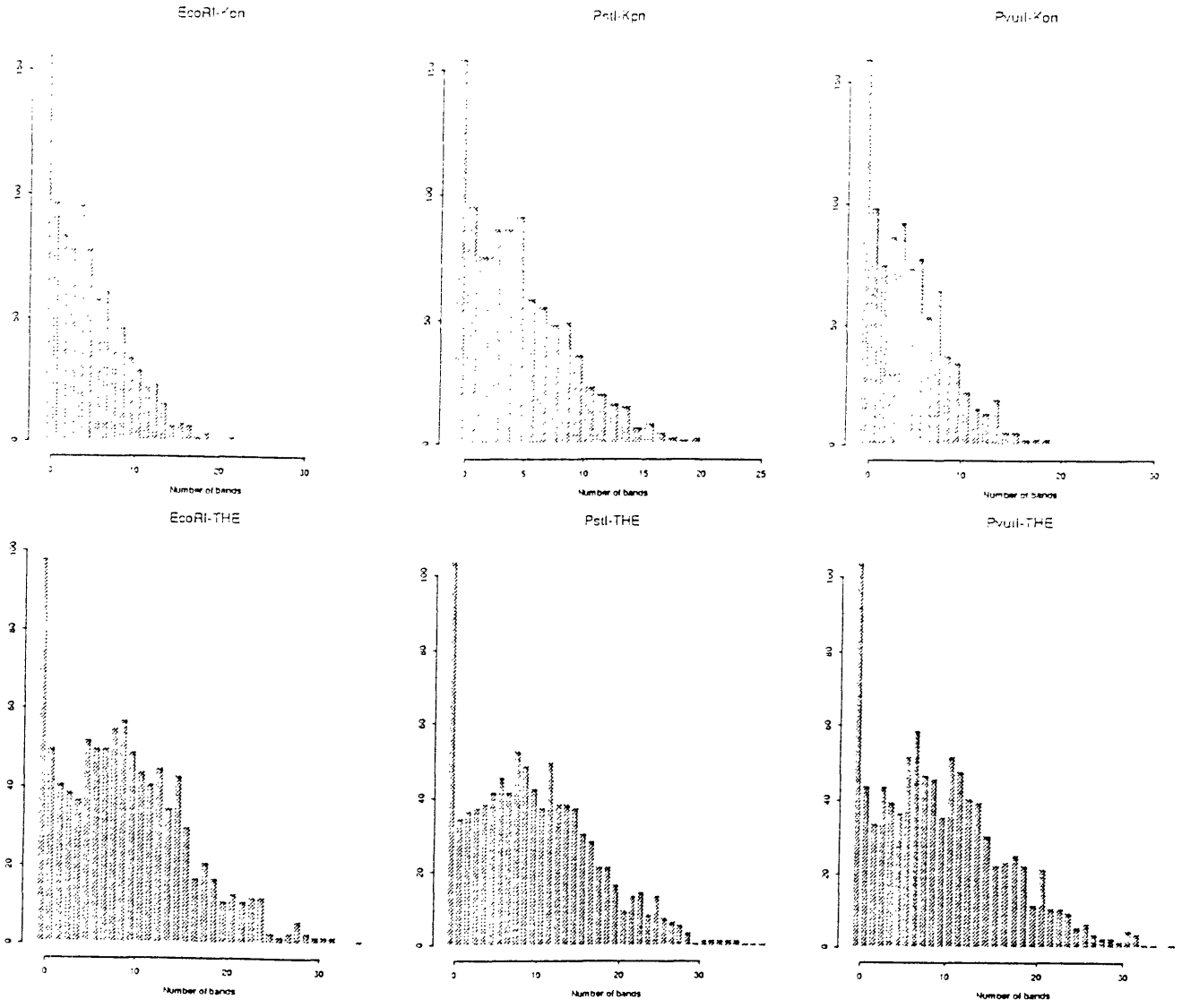


Figure B-14: Number of Band Histograms

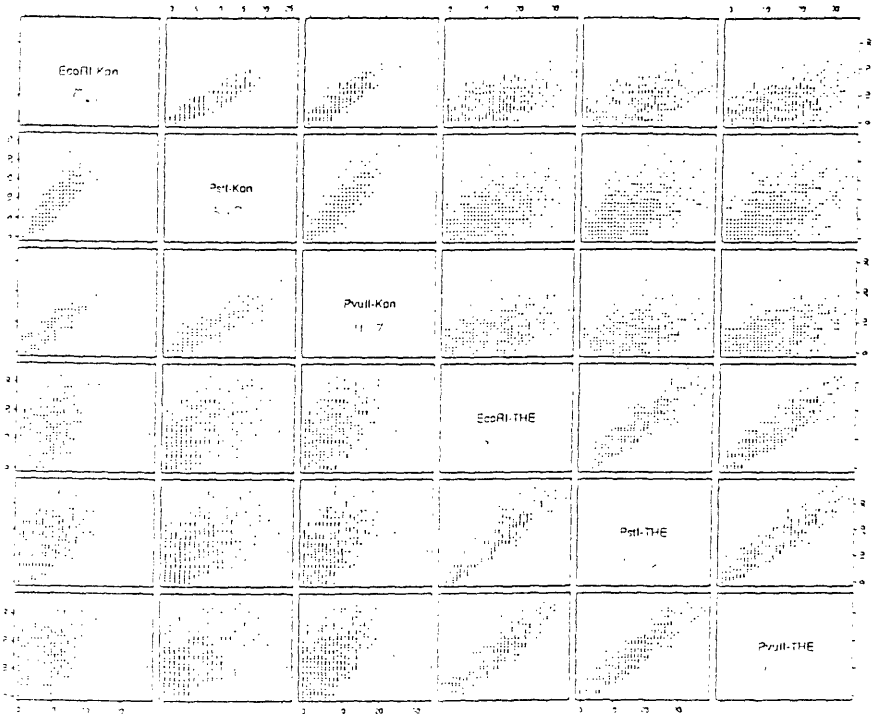


Figure B-15: Number of Band Scatter-plots

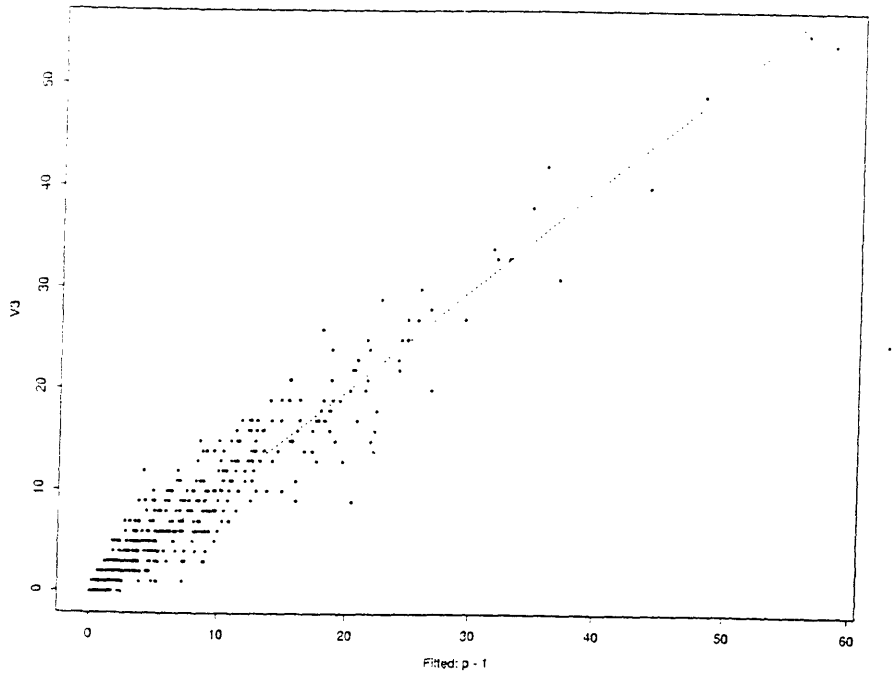


Figure B-16: Justifying $q_i = p_{gain} x_i y_i$ with Linear Regression

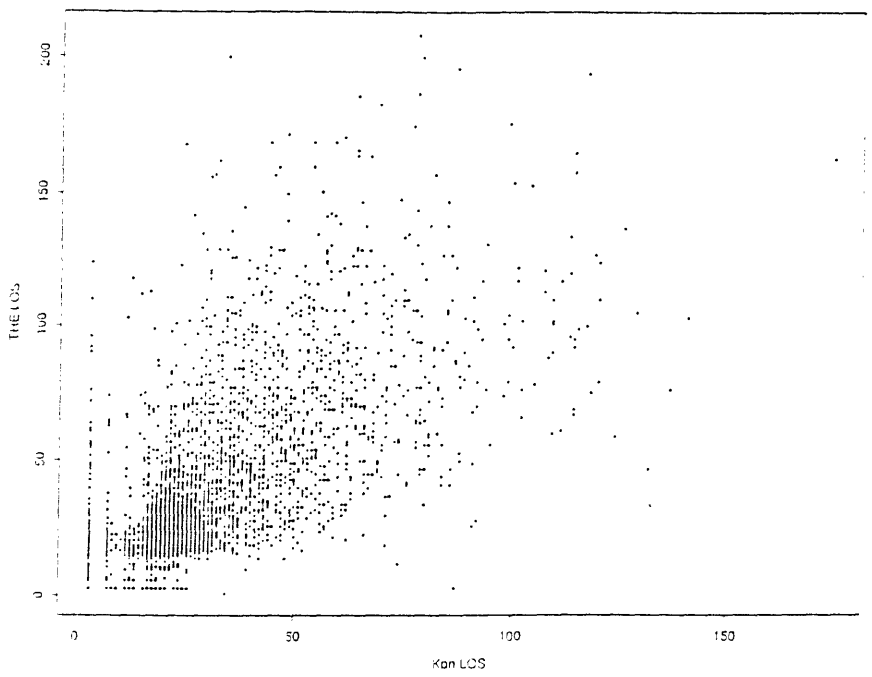


Figure B-17: Correlation between THE and KPN

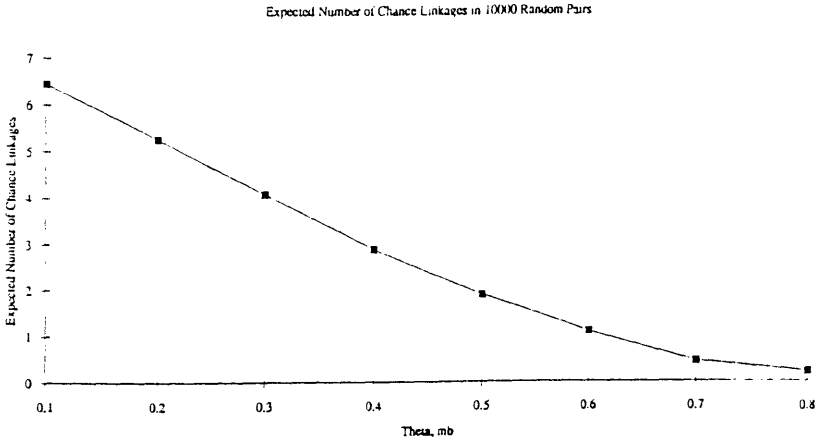


Figure B-18: Matches by Chance in 10000 Random Pairs

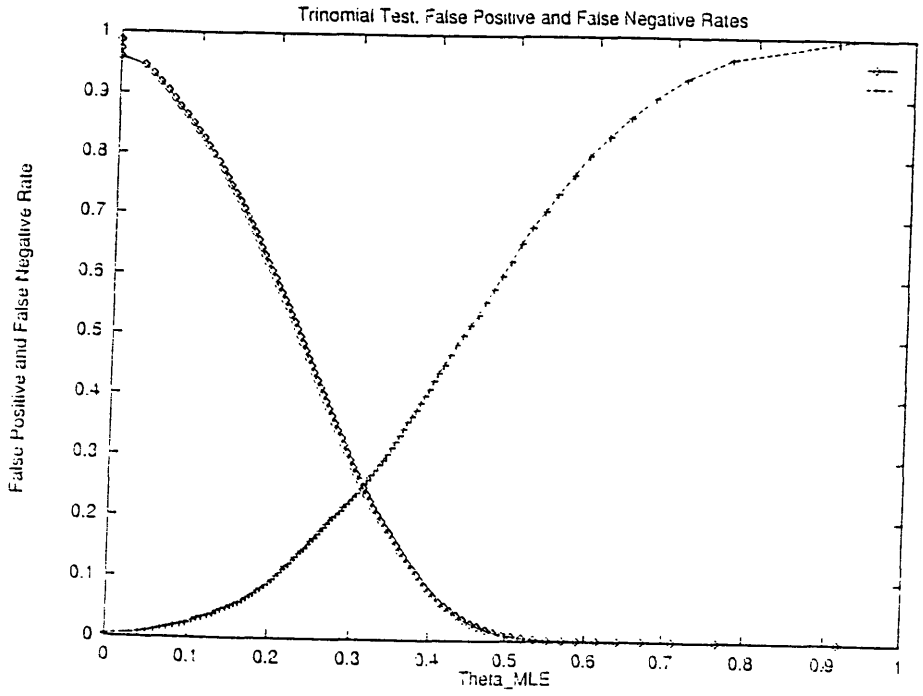


Figure B-19: False Negative and False Positive Rates, `Trinomial` Test

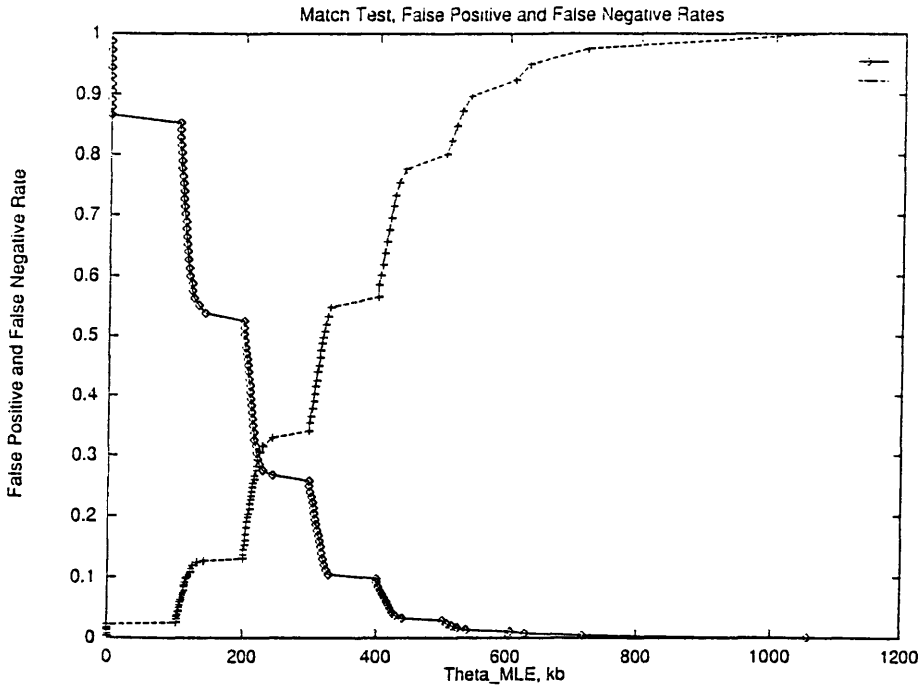


Figure B-20: False Negative and False Positive Rates, `Match` Test

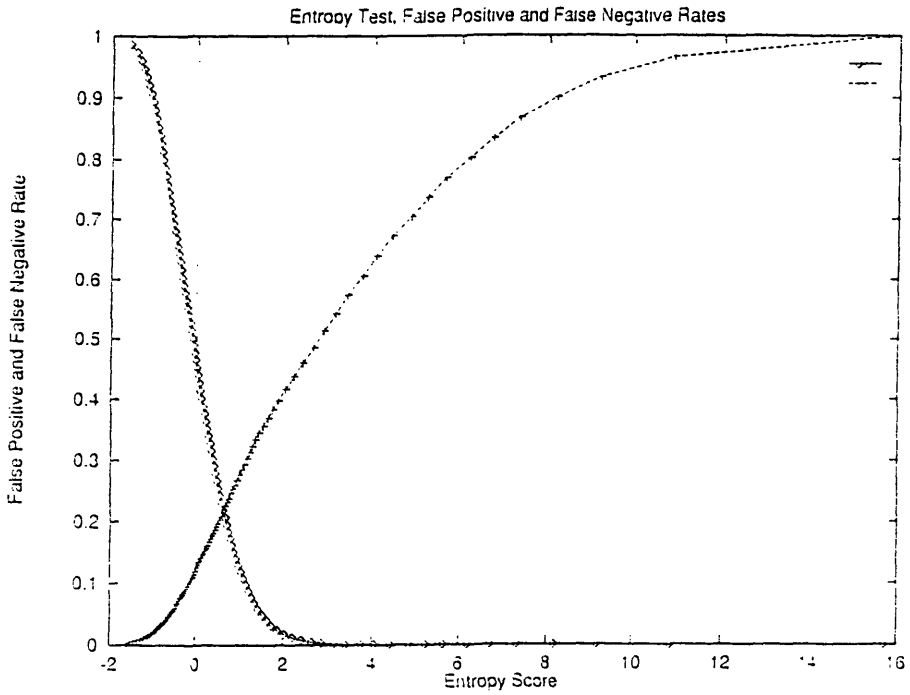


Figure B-21: False Negative and False Positive Rates. Entropy Test

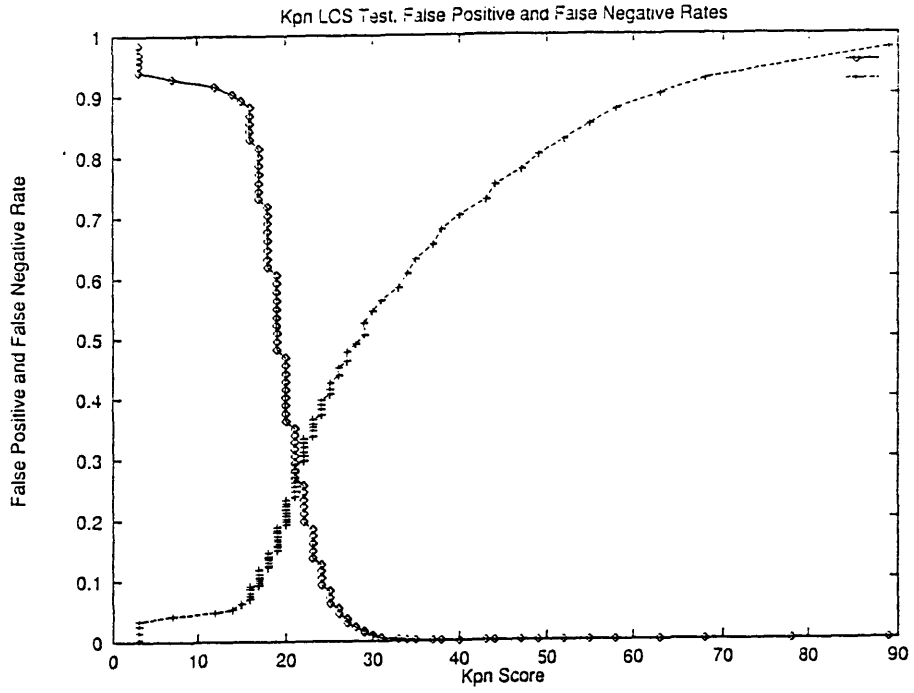


Figure B-22: False Negative and False Positive Rates, KPN Test

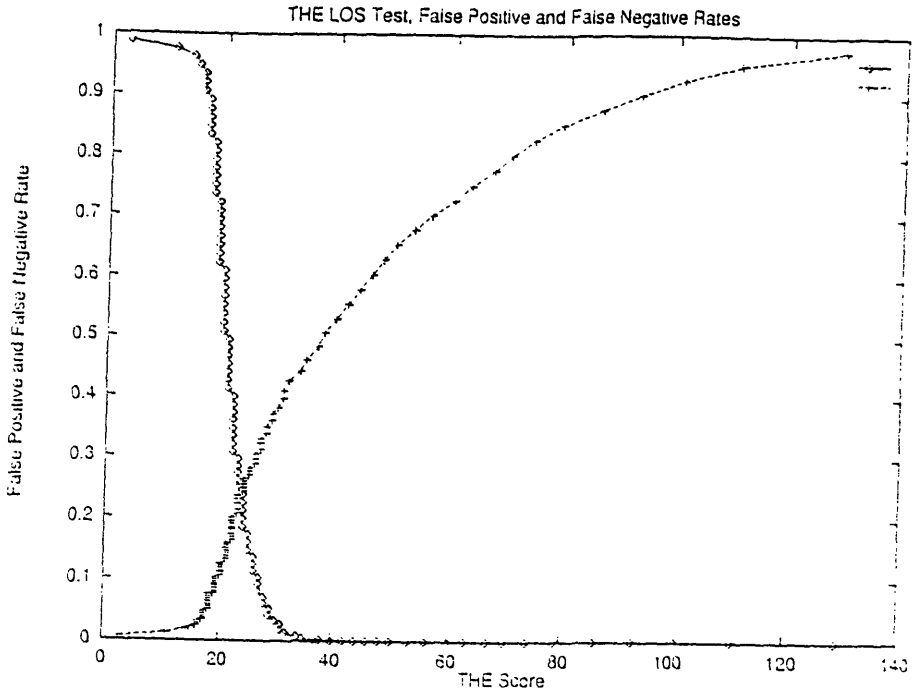


Figure B-23: False Negative and False Positive Rates, THE Test

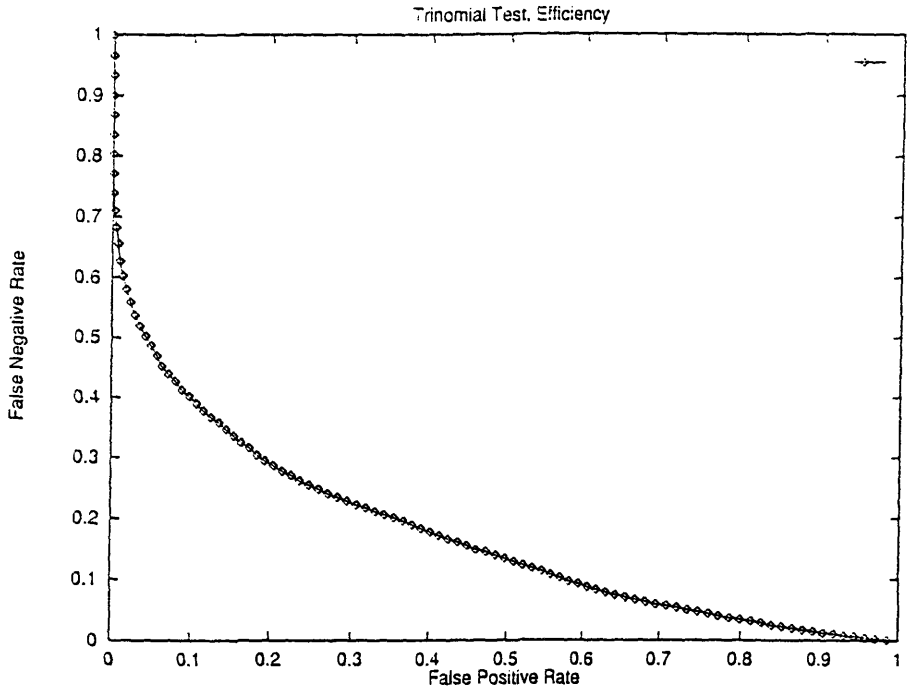


Figure B-24: Efficiency, Trinomial Test

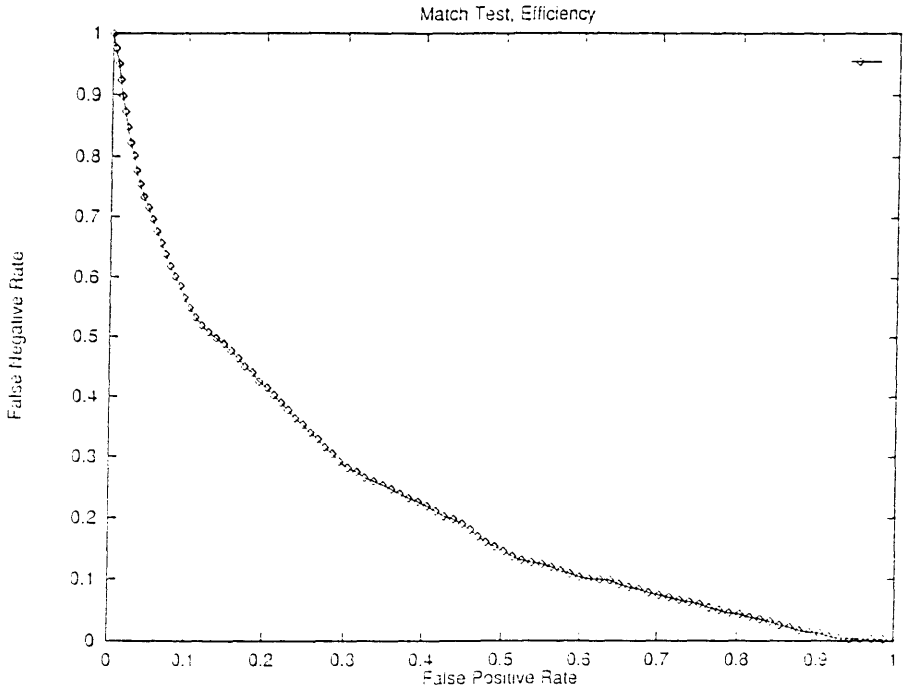


Figure B-25: Efficiency, Match Test

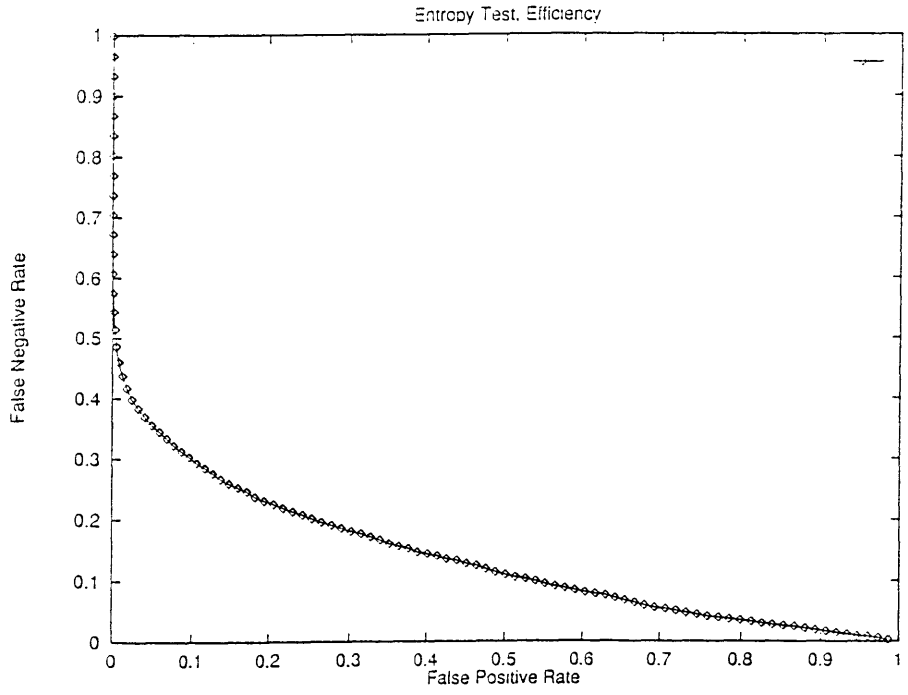


Figure B-26: Efficiency, Entropy Test

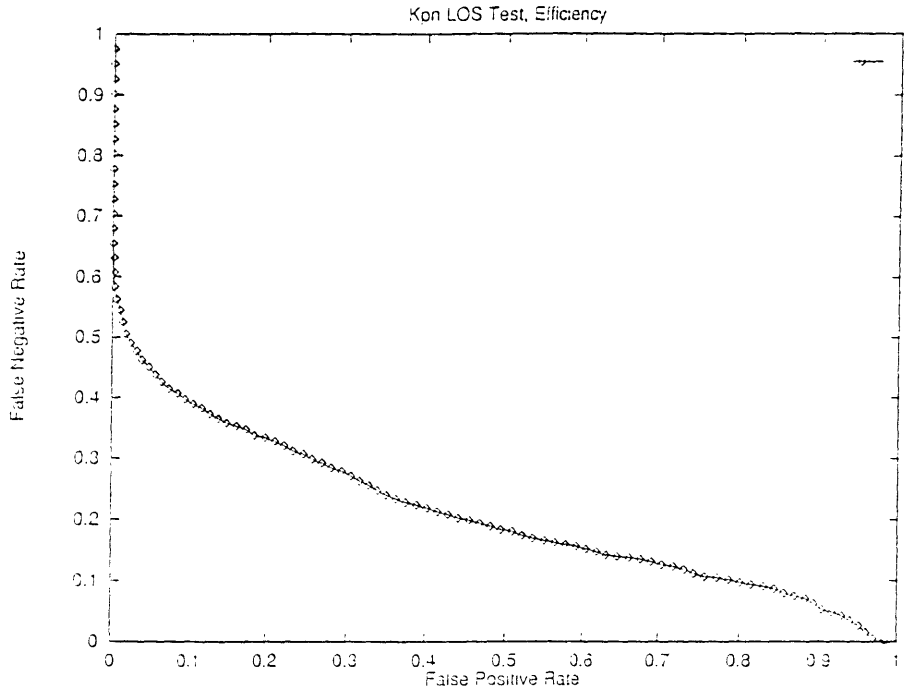


Figure B-27: Efficiency, KPN Test

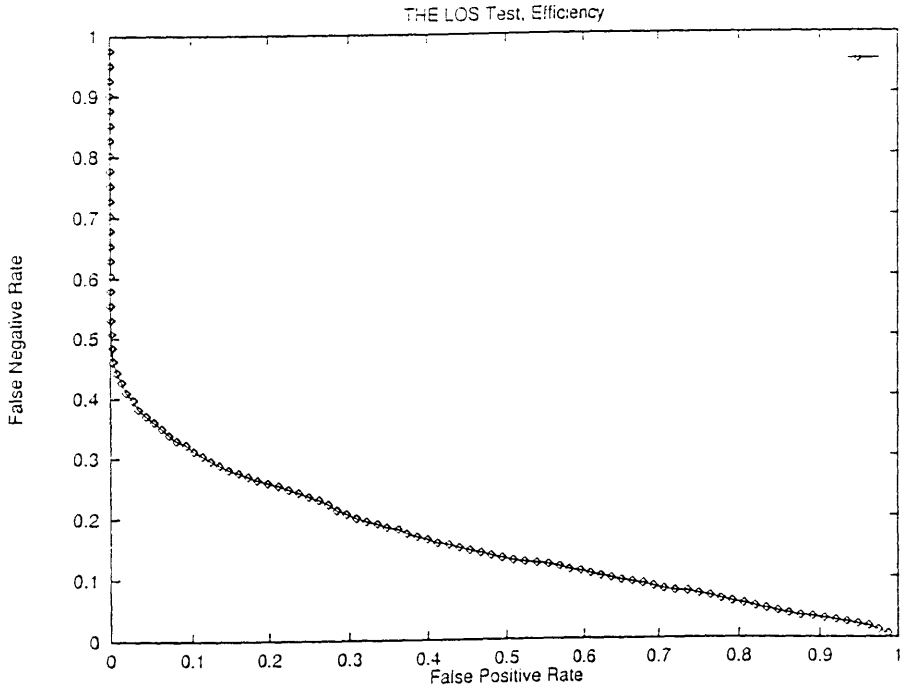


Figure B-28: Efficiency, THE Test

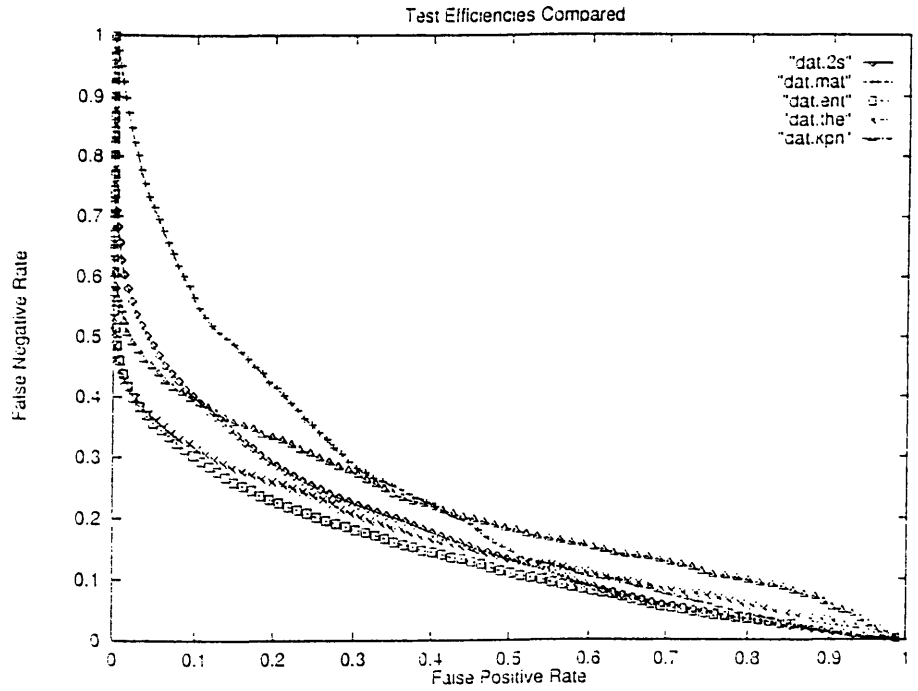


Figure B-29: Test Efficiencies Compared

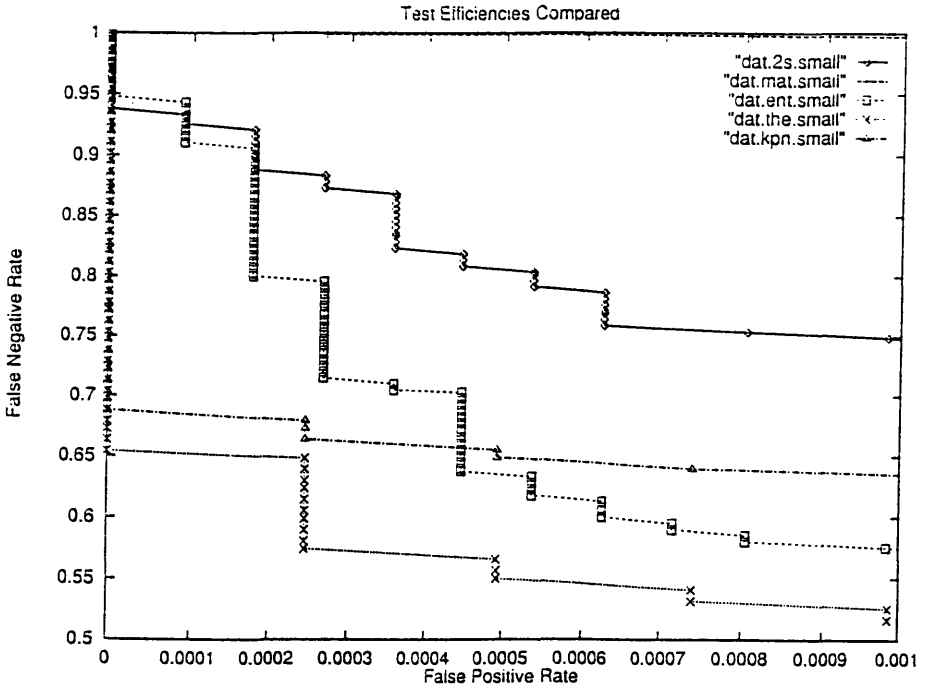


Figure B-30: Test Efficiencies Compared, Small False Positive Rate

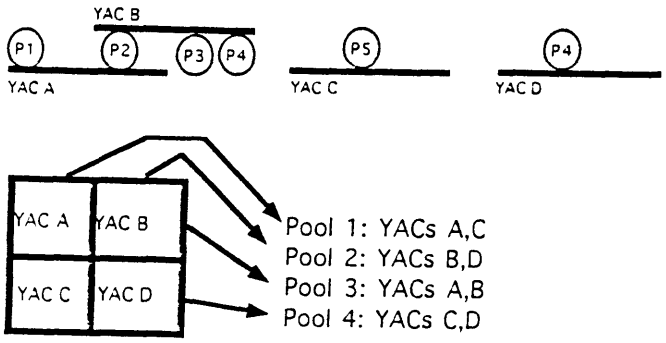


Figure B-31: ALU Probe Screening Example

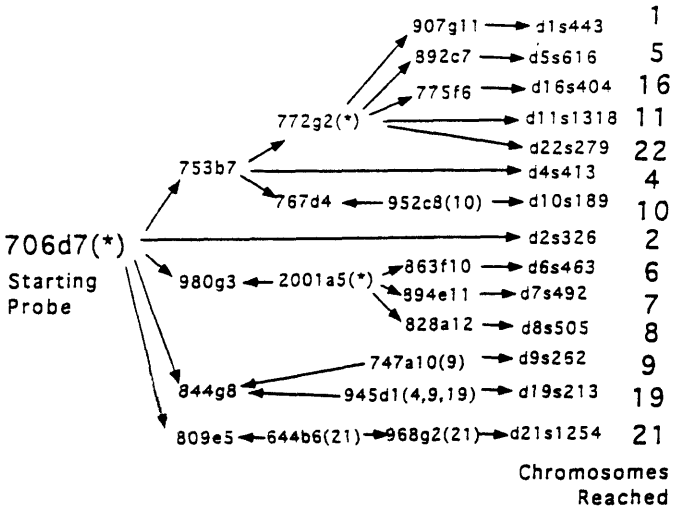


Figure B-32: A Spurious Tree

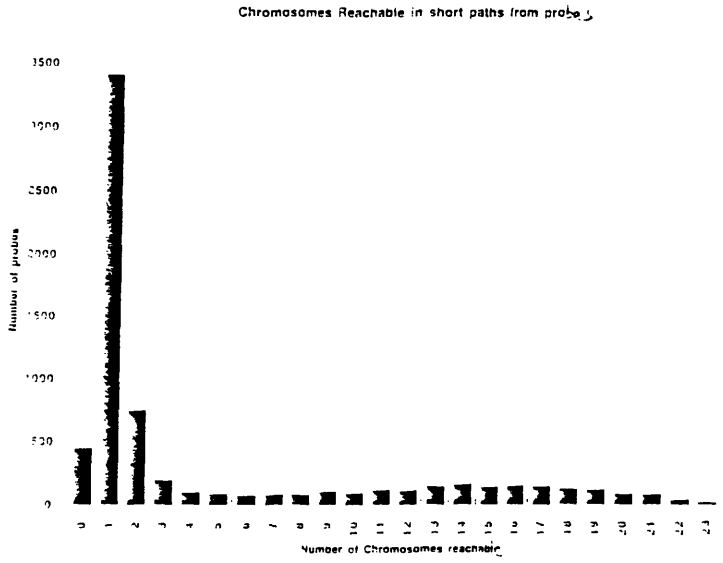


Figure B-33: Chromosomes Reached With Short Paths from Probes

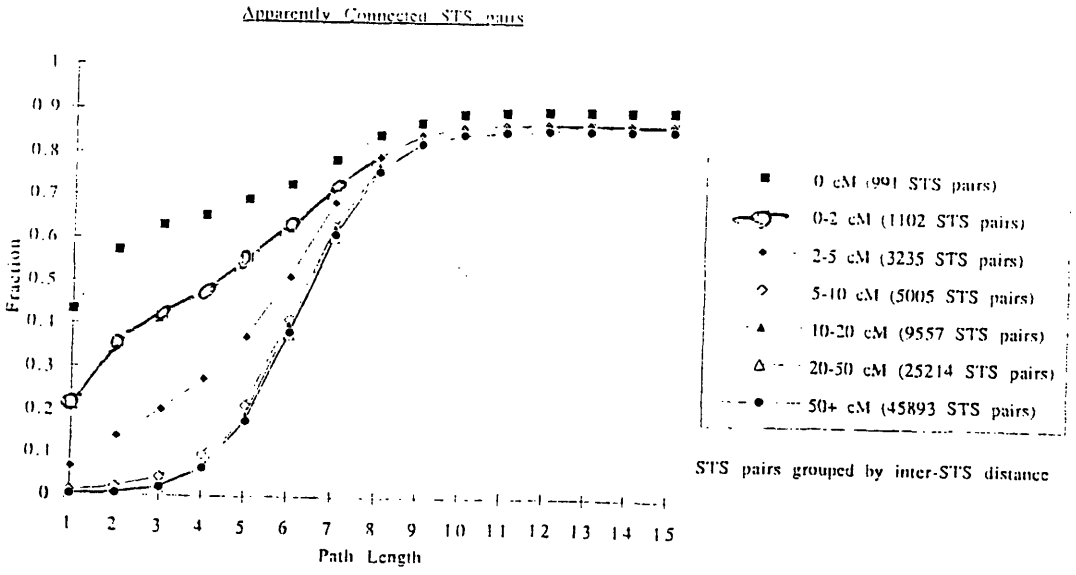


Figure B-34: Fraction of Connected STS Pairs

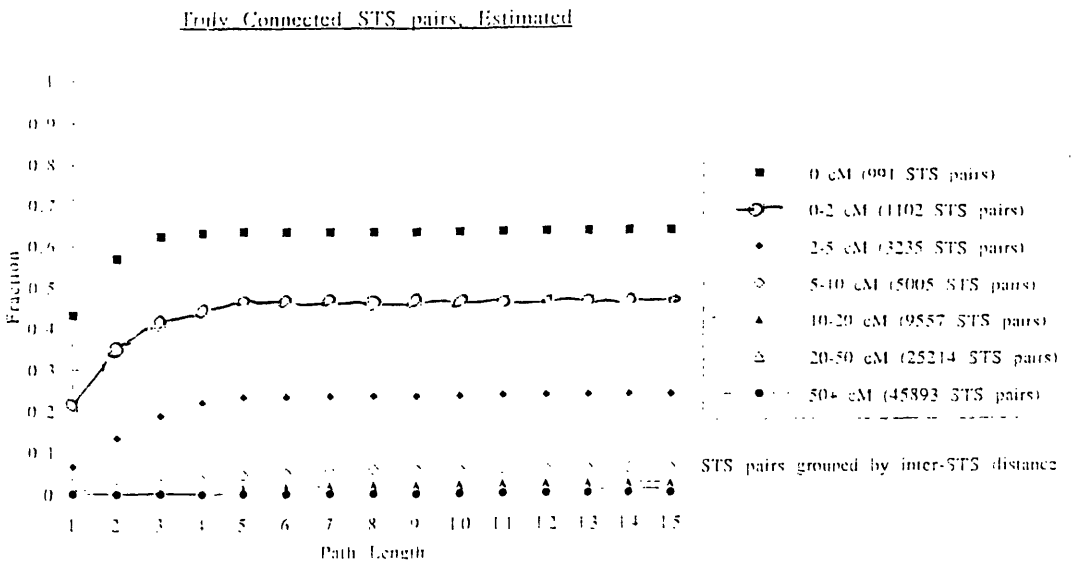


Figure B-35: Fraction of Truly Connected STS Pairs

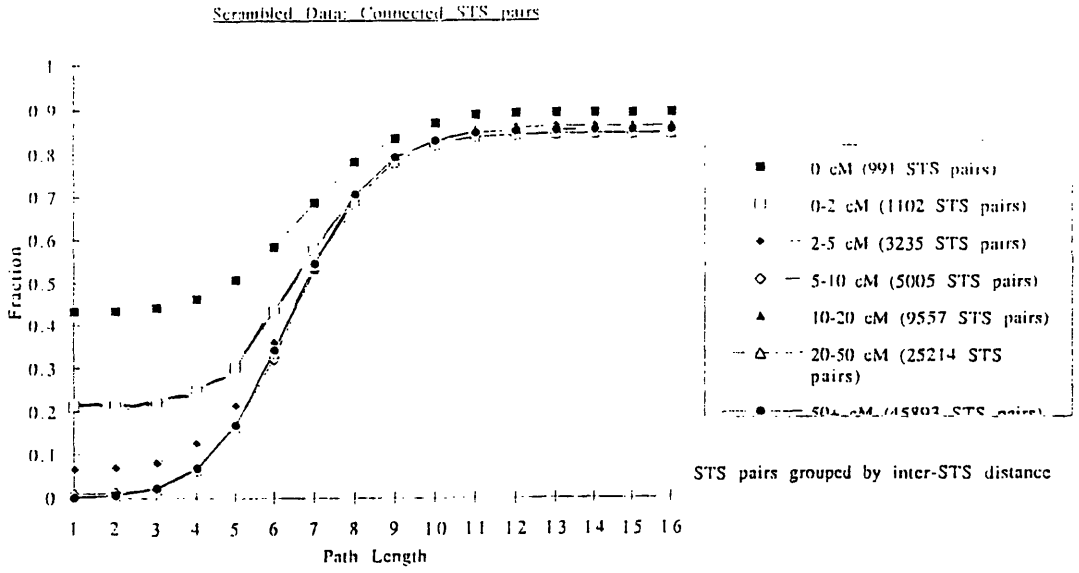
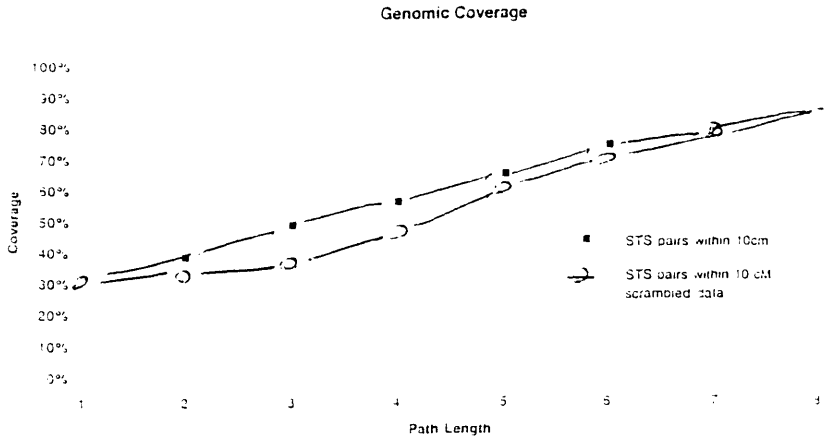


Figure B-36: Fraction of Connected STS Pairs, Scrambled Data



Page 1

Figure B-37: Genome Coverage Using CEPH-Genethon Rules. Real and Scrambled

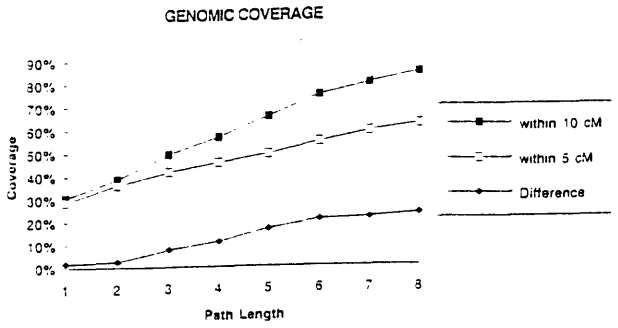


Figure B-38: Genome Coverage Using `Within10cM` and `Within5cM`

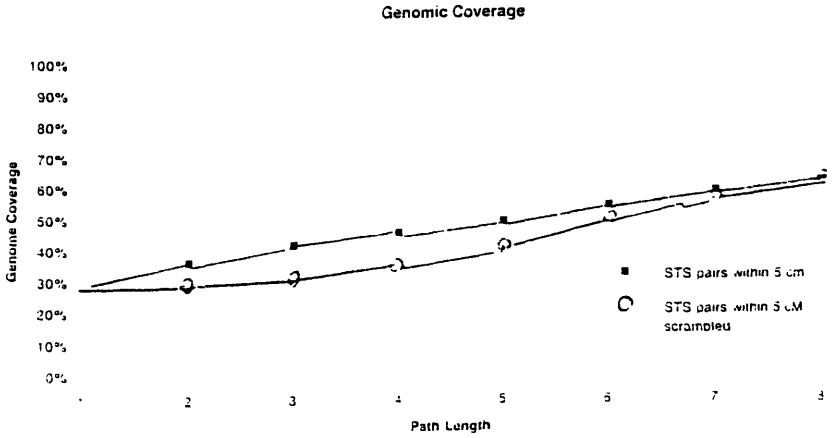


Figure B-39: Genome Coverage Using `Within5cM` Real and Scrambled

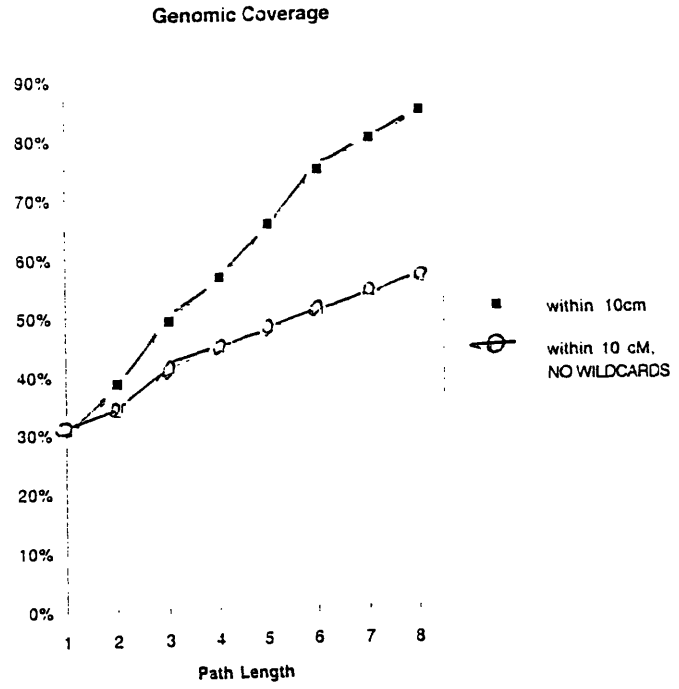


Figure B-40: Genome Coverage Using `UseWildcardProbes` and `NoWildcardProbes`

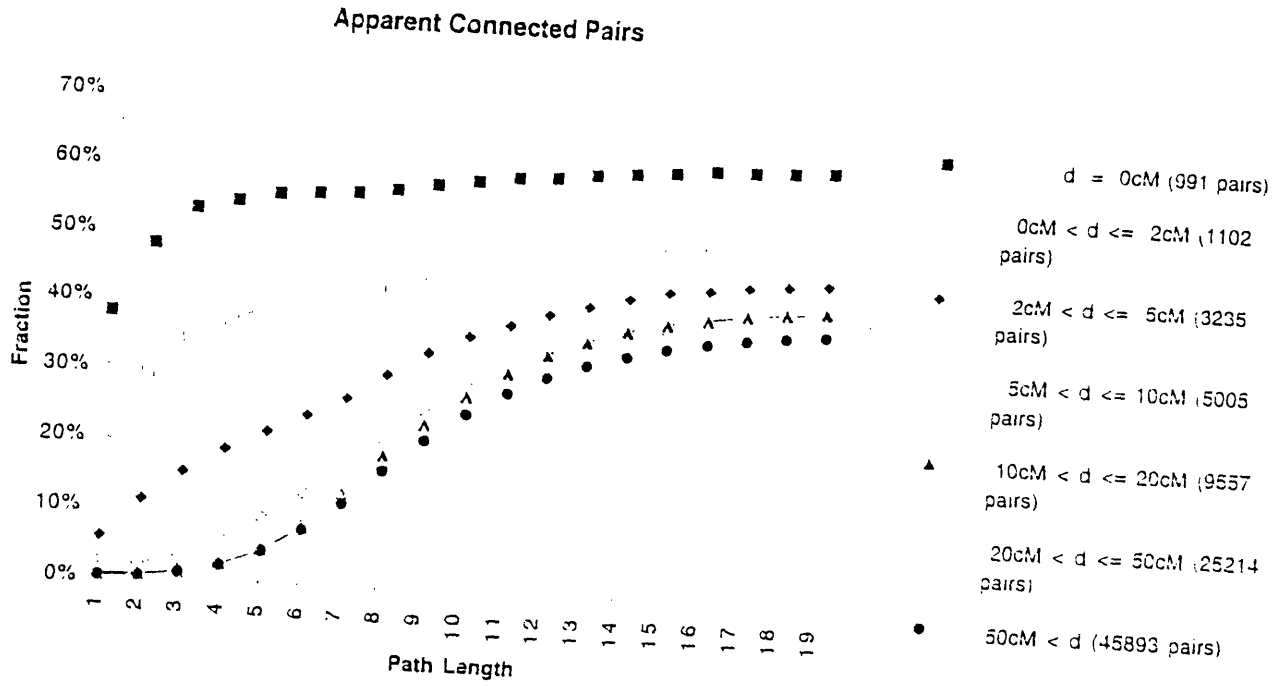


Figure B-41: Fraction of Connected STS pairs Using `NoWildcardProbes`

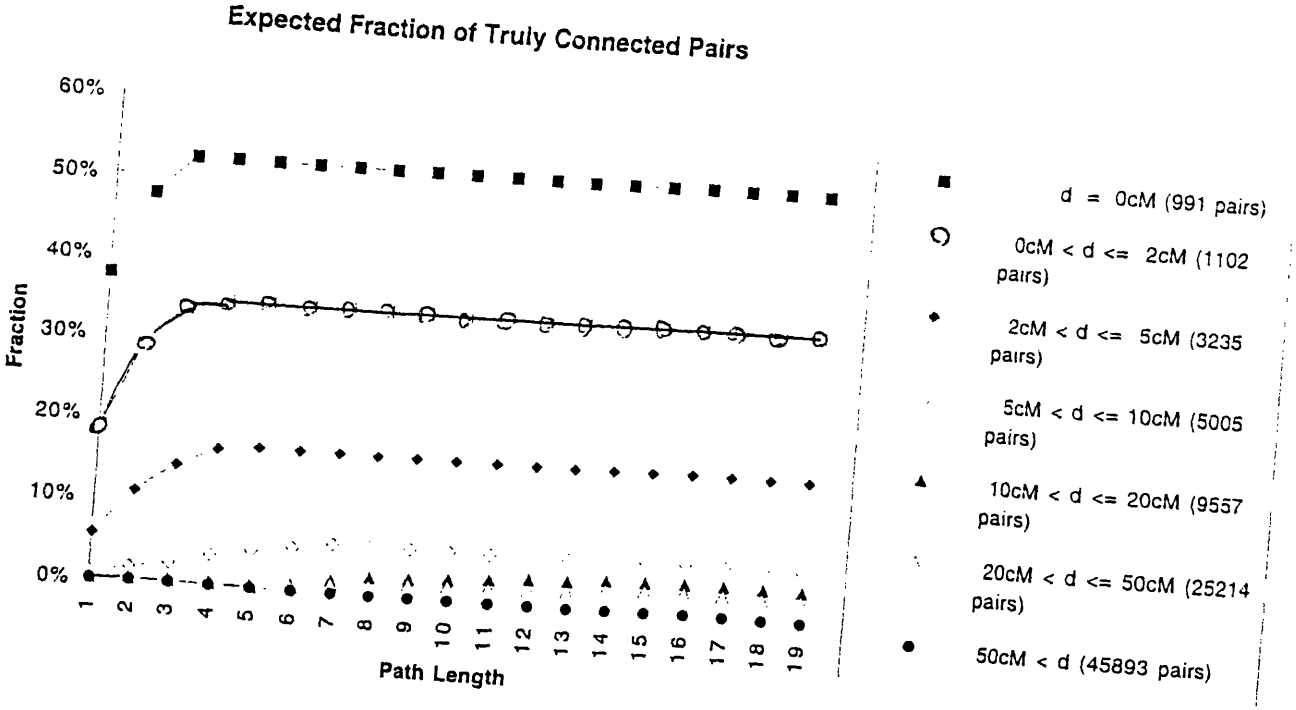


Figure B-42: Fraction of Truly Connected STS Pairs Using `NoWildcardProbes`

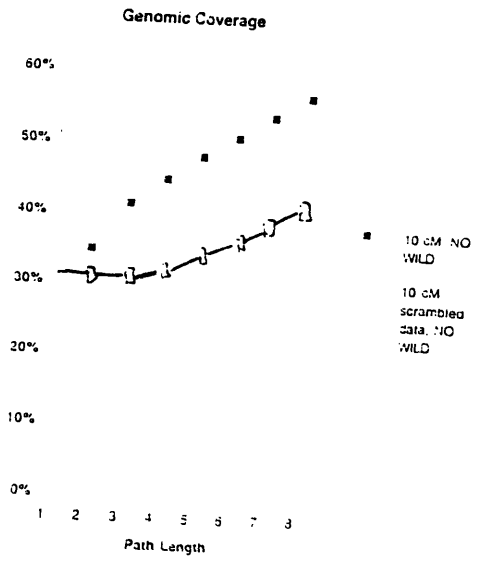


Figure B-43: Genomic Coverage Using `NoWildcardProbes`, Real and Scrambled Data

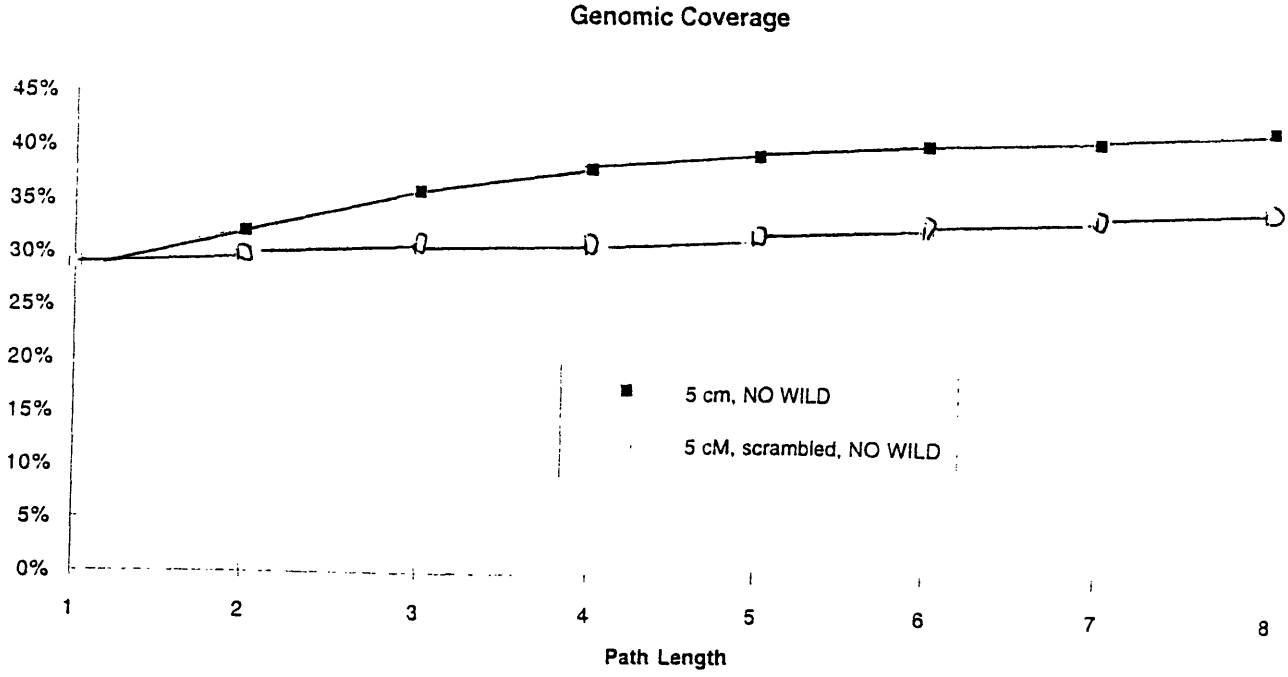


Figure B-44: Genomic Coverage Using `NoWildcardProbes` and `Within5cM`, Real and Scrambled Data

Bibliography

[1] Ravindra Ahuja, Thomas Magnanti, and James Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, Inc., NJ, 1993.

[2] Farid Alizadeh, Richard Karp, Lee Newberg, and Deborah Weisser. Physical mapping of chromosomes: a combinatorial problem in molecular biology. Technical report, Computer Science Division, University of California, Berkeley, September 1992.

[3] Farid Alizadeh, Richard Karp, Deborah Weisser, and Geoffrey Zweig. Physical mapping of chromosomes using unique probes. Technical report, Computer Science Division, University of California, Berkeley, 1994.

[4] Richard Arratia, Eric Lander, Simon Tavare, and Michael Waterman. Genomic mapping by anchoring random clones: a mathematical analysis. *Genomics*, 11:807–827, 1991.

[5] David Balding and David Torney. Statistical analysis of DNA fingerprint data for ordered clone physical mapping of human chromosomes. *Bulletin of Mathematical Biology*, 53:853–879, 1991.

[6] David Baltimore et al. DNA sequencing. *Los Alamos Science*, 20:151–159, 1992.

[7] Sandro Banfi, Susan Ledbetter, A. Chinault, and Huda Zoghbi. An easy and rapid method for the detection of chimeric yeast artificial chromosomes. *Nucleic Acids Research*, 20(7):1814, 1992.

[8] Christine Bellane-Chantelot, Bruno LaCroix, Pierre Ougen, et al. Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell*, 70:1059–1068, 1992.

[9] C. Bellanne-Chantelot, E. Barillot, B. Lacroix, D. Le Paslier, and D. Cohen. A test case for physical mapping of the human genome by repetitive sequence fingerprints: construction of a physical map of 420 kb yac subcloned into cosmids. *Nucleic Acids Research*, 19(3):505–510, 1991.

[10] Béla Bollobás. *Random Graphs*. Academic Press, Harcourt Brace Jovanovich, New York, 1985.

[11] Béla Bollobás and Andrew Thomason. Random graphs of small order. In M. Karoński and A. Ruciński, editors, *Random Graphs 1983*, pages 47–97. North-Holland Mathematical Studies, 1985. 118.

[12] Kellogg Booth and George Lueker. Testing for the consequtive ones property, interval graphs, and graph planarity using PQ tree algorithms. *Journal of Computer and System Sciences*, 13:335–379, 1976.

[13] Ilya Chumakov et al. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature*, 359:380–387, October 1992.

[14] Ilya Chumakov et al. Isolation of chromosome 21-specific yeast articficial chromosomes from a total human genome library. *Nature Genetics*, 1:222–225, June 1992.

[15] D. Cohen, I. Chumakov, and J. Weissenbach. A first-generatiopn physical map of the human genome. *Nature*, 366:698–671, December 16 1993.

[16] Necia Grant Cooper. The human genome project. *Los Alamos Science*, 20:1–337, 1992.

[17] Alan Coulson, John Sulston, Sydney Brenner, and Jonathan Karn. Toward a physical map of the genome of the nematode Caenorhabditis Elegans. *Proceedings of the National Academy of Sciences USA*, 83:7821–7825, October 1986.

[18] David Cox, Margit Burmesiter, E. Roydon Price, Suwon Kim, and Richard Meyers. Radiation hybrid mapping: A somatic cell genetic method for constructing high resolution maps of mammalian chromosomes. *Science*, 250:245–250, October 1990.

[19] Sougata Datta. Ceph fingerprints and their analysis. Master's thesis, Massachusetts Institute of Technology, Cambridge. Massachusetts, 1994. Operations Research Center.

[20] P. J. DeJong et al. Human chromosome-specific partial digest libraries in λ and cosmid vectors. In *Human Gene Mapping 10*, volume 51, page 985, New Haven conference, 1989. Cytogenics and Cell Genetics.

[21] Alvin W. Drake. *Fundamentals of Applied Probability Theory*, chapter 4, pages 123–153. McGraw-Hill Publishing Company, 1967.

[22] James Fickett and Michael Cinkosky. A genetic algorithm for assembling chromosome physical maps. In H. Lim, J. Fickett, C. Cantor, and R. Robbins, editors, *The Second International Conference of Bioinformatics, Supercomputing, and Complex Genomic Analysis*, pages 273–285, New Jersey, 1992. World Scientific.

[23] Simon Foote. Personal Communication, January 1994.

[24] Simon Foote, Douglas Vollrath, Adrienne Hilton, and David Page. The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science*, 258:60–66, October 2 1992.

[25] M. Golumbic, H. Kaplan, and R. Shamir. On the complexity of dna physical mapping. Technical Report 271, The Moise and Frida Eskenasy Institute of Computer Sciences, Tel Aviv University, Israel, January 1993.

[26] David Greenberg and Sorin Istrail. Algorithmic analysis of physical mapping in the presence of chimeric clones: Progress report. Technical report, Sandia National Laboratory, Algorithms and Discrete Mathematics, May 22 1993.

[27] John Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, Michigan, 1975.

[28] Tom Hudson. Personal Communication, 1994.

[29] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 13 1983.

[30] Yuji Kohara, Kiyotaka Akiyama, and Katsumi Isono. The physical map of the whole E. Coli chromosome: Application of a new strategy for rapid analysis and sort of a large genomic library. *Cell*, 50:495–508, July 31 1987.

[31] Bruno LaCroix. Personal Communication, 1994.

[32] Bruno Lacroix and Jean-Jacques Codani. Computational aspects of human genome physical mapping. Technical Report 1560, INRIA, France, 1991.

[33] Bruno Lacroix and Jean-Jacques Codani. Physical mapping of the human genome: computational aspects. Technical report, CEPHB-Genethon, INRIA, Paris, France, 1992.

[34] Eric S. Lander and Michael S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2:231–239, 1988.

[35] Richard Larsen and Morris Marx. *An Introduction to Mathematical Statistics and Its Application*, chapter 5, pages 259–287. Prentice-Hall, Englewood Cliffs, New Jersey, 2 edition, 1986.

[36] Richard Larson and Amadeo Odoni. *Urban Operations Research*, chapter 3, pages 77–162. Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[37] Eugene Meyers. Advances in sequence assembly. Technical report, Department of Computer Science, University of Arizona, Tucson, Arizona, 1992.

[38] Richard Mott and Andrei Grigoriev. Programs for analysing hybridization data. Technical report, Genome Analysis Laboratory, Imperial Cancer Research Fund, October 26 1992.

[39] R. A. Mulivor et al. Characterization of DNA from single human chromosome hybrids. In *American Journal of human genetics, Supplement*, volume 49, page 370, 1991.

[40] Lee Newberg. Finding a most likely clone ordering from oligonucleotide hybridization data. Technical report, Biological Sciences Division, University of Chicago, Chicago, Illinois, December 1993.

[41] Lee Newberg. *Finding, evaluating, and counting contig maps.* PhD thesis, University of California at Berkeley, Department of Computer Science, 1993.

[42] Maynard Olson, Leroy Hood, Charles Cantor, and David Botstein. A common language for physical mapping of the human genome. *Science*, 245:1434–1440, September 29 1989.

[43] Maynard V. Olson et al. Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences USA*, 83:7826–7830, October 1986.

[44] DOE Human Genome Program. Primer on molecular genetics. Office of Health and Environmental Research, Office of Energy Research, U. S. Department of Energy, April 1992.

[45] Mary Pat Reeve, Mark Daly, Alan Kaufman, Stephen Lincoln, Simon Foote, James Orlin, Eric Lander, and Nat Goodman. A contig assembly algorithm for mapping the human genome. Slide Presentation, August 1993.

[46] Philippe Rigault. Clone ordering by simulated annealing: application to STS content map of chromosome 21. Technical report, Genethon, Paris, France, 1992.

[47] Philippe Rigault. personal communication, April 1994. Cold Springs Conference on Genome Mapping.

[48] Philippe Rigault. QUICKMAP program and QUICKMAP data release. anonymous FTP from CEPH-genethon-map.genethon.fr, March 1994.

[49] Leslie Roberts. Two chromosomes down, 22 to go. *Science*, 258:28–30, October 2 1992. Research News.

[50] C. Soderlund and C. Burks. GRAM and GENEFRAGII: Solving and testing the single digest partially-ordered restriction map problem. To appear in CABIOS.

[51] C. Soderlund, D. Torney, and C. Burks. Calculating shared fragments for the single digest problem. In *Proceedings of the 26th Hawaii International Conference on Systems Science*, volume 1, pages 620–629. IEEE Computer Society Press, 1993.

[52] Raymond Stallings, David Torney, Carl Hildebrand, Jonathan Longmire, Larry Deaven, James Hett, Norman Doggett, and Robert Moyzis. Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proceedings of the National Academy of Sciences USA*, 87:6218–6222, 1990.

[53] John Sulston, Frank Mallet, Rodger Staden, Richard Durbin, Terry Hornsell, and Alan Coulson. Software for genome mapping by fingerprinting techniques. *CABIOS*, 4(1):125–132, 1988.

[54] David Torney, Clive Whittaker, Steven White, and Karen Schenk. Computational methods for physical mapping of chromosomes. In *Proceedings of the conference on electrophoresis, supercomputing, and the human genome*, Tallahassee, Florida, April 10-13 1990. Florida State University.

[55] E. G. Valiant. Complexity of computing the permanent. *Theoretical Computer Science*, 8:189–201, 1979.

[56] Jean Weissenbach, Gabor Gyapay, Colette Dib, Alain Vignal, Jean Morissette, Phillipe Millasseau, Guy Vaysseix, and Mark Lathrop. A second-generation linkage map of the human genome. *Nature*, 359:794–801, October 29 1992.

[57] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*, chapter 12, pages 466–472. Springer Series in Statistics. Springer-Verlag, 1989.

[58] Clive C. Whittaker, Mark Mundt, Vance Faber, David Balding, Randall Dougherty, Raymond Stallings, Steven White, and David Torney. Computations for mapping genomes with clones. *International Journal of Genome Research*, 1(3):195–226, 1993.