# The distribution of restriction enzyme sites in *Escherichia coli*

Gary A.Churchill, Donna L.Daniels[1] and Michael S.Waterman
Department of Mathematics, University of Southern California. Los Angeles, CA90089 and [1]Laboratory of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received September 29, 1989; Revised and Accepted December 28, 1989

## ABSTRACT

**A statistical analysis of physical map data for eight restriction enzymes covering nearly the entire genome of *E. coli* is presented. The methods of analysis are based on a top-down modeling approach which requires no knowledge of the statistical properties of the base sequence. For most enzymes, the distribution of mapped sites is found to be fairly homogeneous. Some heterogeneity in the distribution of sites is observed for the enzymes *Pst*I and *Hin*dIII. In addition, *Bam*HI sites are found to be more evenly dispersed than we would expect for random placement and we speculate on a possible mechanism. A consistent departure from a uniform distribution, observed for each of the eight enzymes, is found to be due to a lack of closely spaced sites. We conclude from our analysis that this departure can be accounted for by deficiencies in the physical map data rather than non-random placement of actual restriction sites. Estimates of the numbers of sites missing from the map are given, based both on the map data itself and on the site frequencies in a sample of sequenced *E. coli* DNA. We conclude that 5 to 15% of the mapped sites represent multiple sites in the DNA sequence.**

## INTRODUCTION

The genome of *E. coli* is a 4.72Mb circular dsDNA molecule. It has been extensively mapped by both genetic and physical methods and may soon be the first large genome to be completely sequenced. The existence of a detailed restriction map containing over 7000 sites for eight different restriction enzymes and covering nearly the entire genome (1), affords us with a unique opportunity to examine the distribution of these sites in detail. Such analysis is aimed at identifying interesting structural or organizational features of the genome or systematic bias in the data. A detailed understanding of the relationship between physical map data and the actual placement of sites in a genome will aid the process of integrating physical and genetic maps with sequence data (2). Models for restriction site distributions could also have practical applications in the early planning stages of a mapping or sequencing project where the choice of an appropriate set of restriction enzymes would help to minimize the amount of work required. Results presented here for the *E. coli* genome will very likely generalize to other bacterial species, although, in higher organisms, the distribution of sites could be very different due the fundamentally different organization of genomic DNA.

The hypothesis that the restriction sites are uniformly and independently distributed throughout the genome is to be tested. There are several *a priori* reasons to suspect that this will not be the case. First, if there are local fluctuations in base composition, regions will exist where the expected frequency of sites may be increased or decreased relative to the average. There is strong evidence for compositional fluctuations in vertebrate DNA (3). Density gradient analysis of fragmented bacterial genomes also shows variations in excess of random expectations (4). Second, the DNA of *E. coli* is not a random sequence, although we treat it as such for purposes of analysis. There is a high density of genetic information in the genome which is dominated by genes and regulatory regions. Coding constraints and codon usage could have a significant effect on the distribution of sites. Finally, certain specific sequence patterns play important roles in the function of the organism and their placement could be highly non-random. Among the enzyme included in this study, *Eco*RI and *Eco*RV are *E. coli* specific restriction enzymes and the BamHI recognition pattern contains within it the Dam methylation site GATC, which has been shown to play a role in excision repair processes (5). None of the eight enzymes studied here contains the rare CTAG tetramer (6).

The distribution of restriction enzyme sites has been considered by several authors (7–10). It is usually assumed that the bases of the DNA form a homogeneous Markov chain and the process of restriction enzyme cutting is modeled as a Markov renewal process. Exact results obtainable by these methods are unneccessarily complex for our purposes and we expect that, for many genomes, sufficient data on oligonucleotide frequencies will not be available. Our approach is to model the placement of restriction sites directly and thus avoid many of the problems associated with modeling of the base sequence. We will assume that the sites are placed uniformly throughout the genome, independently of the placement of any other sites, i.e. that the sites constitute a Poisson process. Problems of site overlap and neighboring base interactions are ignored. Thus, the model is expected to be inaccurate over very short distances ($<$20bp) but should be sufficient for most practical purposes. An important implication of the Poisson process model is that the distances between adjacent sites will follow an exponential distribution. Bounds on the quality of the Poisson approximation to the Markov renewal process can be obtained using theorem 2 of Arratia et al. (11).

In the methods section below, we provide a brief description of the map data and some relevant experimental aspects of the methods used to construct the map. This is followed by an outline of the statistical methods used in the analysis. Results are presented in four sections. First, we consider the placement of mapped sites and compute statistics which test for uniformity. Next, we examine the fragment length data, including largest and smallest fragments. Then, a refinement of the Poisson process model is proposed by which the map data is generated as an incomplete representation of the true restriction pattern. The modified Poisson process model allows for the fact that, due to the methods used to construct the map, closely spaced sites are frequently missed. Estimates of the actual numbers of sites are obtained. An alternative estimate, based on a sample of sequenced *E. coli* DNA, is computed for comparison. Finally, the alignment of three sequenced fragments to the map confirms the basic premises of the model.

## METHODS

### Data

*E. coli Restriction Map.* For this analysis, we used the nearly complete eight enzyme restriction map of *E. coli* presented by Kohara et al. (1, figure 6). This map was constructed by merging individual maps of each of a large set of overlapping regions of the genome which had been cloned and mapped in phage lambda vectors. The individual maps were experimentally determined by a modified Smith-Birnstiel method (12).

In this method, DNA to be mapped is labeled at a unique end, partially digested with the enzyme to be mapped, electrophoresed, and the labelled fragments are visualized by autoradiography. Deducing the map is conceptually straightforward since measuring the length of every labelled fragment maps a restriction site relative to the unique end.

Several types of map errors occur with this methodology and are relevant to our statistical analysis. Measurement errors are quite large. The unique end used to map inserts in lambda vectors was the lambda right end which is at some distance from the cloned insert being mapped. Thus the fragments being measured are in the ten to thirty kb range. This is a poorly resolved region of electrophoresis gels and the error of these measurements is five to fifteen percent. Intersite distances are obtained by differencing the lengths of the measured fragments. The error on these values is hard to estimate but is probably not as large as the sum of the errors of the measured fragments since these errors (from the same lane of the same gel) are not independent. Estimating length errors in the map is further complicated by the fact that maps of many clones were compiled to produce the final map and these were variably weighted based on such factors as the quality of the gel (Kohara, pers. comm.). For these reasons, we believe that the measurement errors in the Kohara map are not simply proportional to the intersite distances.

Besides these quantitative measurement errors, there are also three types of qualitative errors. The eight different enzymes were mapped in eight adjacent lanes on the gel. The relative order of closely spaced sites for different enzymes is sometimes not correct due to the limited resolution of the gel. Similarly, two closely spaced sites for the same enzyme may be mapped as a single site due to failure of the gel to resolve the two fragments into two bands. Finally, some sites may be entirely missed due to variability of sites in rates of cutting and the inability to distinguish between a clone with no sites and a failed digest. This seems to be fairly rare, probably because of the redundancy of the mapping data.

**Table 1:** A Sample of Sequenced *E. coli* DNA

| Locus | Length | Accesssion | Location |
|---|---|---|---|
| ECOACE | 7740 | V01498 | 2.8 |
| ECOAMPCFR | 5482 | J01611 | 94.4 |
| ECOBGLO | 5270 | M16487 | 83.4 |
| ECOBIO | 5793 | J04423 | 18 |
| ECOCARAB | 5227 | J01597 | 0.8 |
| ECODMS | 6492 | J03412 | 20 |
| ECOGLTA | 13063 | J01619 | 16.5 |
| ECOGLTB | 6292 | M18747 | 69.4 |
| ECOHISPUR | 6172 | J02800 | 50.2 |
| ECOHLY | 8211 | M10133 | |
| ECOILVGE | 9456 | M10313 | 85 |
| ECOLAC | 7477 | J01636 | 7.9 |
| ECOLPXA | 6627 | M19334 | 4 |
| ECOMALB | 6545 | J01648 | 91.5 |
| ECONRDA | 8554 | K02672 | 48.5 |
| ECONUSA | 5423 | X00513 | 68.9 |
| ECOPHOS | 5032 | K01992 | 83.6 |
| ECOPURLA | 5865 | M19501 | 55 |
| ECORBS | 5820 | M13169 | 84 |
| ECORECC | 6000 | X03966 | 61 |
| ECORGNB | 7508 | J01695 | 89.8 |
| ECORPLN | 5922 | X01563 | 72.9 |
| ECORPLRPO | 12337 | J01678 | 89.9 |
| ECORPOS10 | 5422 | X02613 | 73 |
| ECORPSRPO | 5059 | J01687 | 67.0 |
| ECOTGP | 7335 | J01714 | 27.5 |
| ECOTHR | 5922 | J01706 | 0.0 |
| ECOTHRINF | 7784 | V00291 | 37.6 |
| ECOUHP | 5400 | M17102 | 82.2 |
| ECOUNCC | 14526 | X01631 | 83.9 |

Sequences were obtained from GenBank release 60. The locus name, length in base pairs, accession number and approximate location on the 100 minute K-12 genetic map are shown.

*Electronic Version of the Kohara Map.* An enlarged version of figure six (1) was obtained from Kohara. This was used as a template for drawing a restriction map using a Macintosh computer and a program supplied by DNAstar inc. After entry, the map was printed to the same scale as the Kohara figure. The two were physically superimposed, the differences located by eye and site positions were adjusted. Data were recorded to the nearest base, although the true accuracy may be less than 0.5kb. Some periodicities in the lengths of short fragments are apparent in the recorded values and are an artifact of the entry method.

*Sampled Sequences.* For purposes of inferring the numbers of restriction sites, a sample of sequenced *E. coli* DNA was selected from the GenBank database. The sample includes all sequenced regions ≥ 5kb in length from GenBank release 60. Duplications were eliminated, leaving 30 sequences with a total length of 213,756bp. Identifying information for these fragments is listed in table 1.

### Statistical Methods

*Notation.* For a given enzyme, we will let $t_1 < \ldots < t_n$ denote the ordered locations of the $n$ sites relative to a fixed origin. We will also consider the scaled sites $u_i = t_i/G$, where $G$, the total genome size, is taken to be 4720kb. The fragment lengths, defined as the distances between two adjacent sites, will be denoted by $x_i = t_i - t_{i-1}$, where $t_0 \equiv t_n$ is used to compute the length of the fragment containing the origin. The ordered fragment lengths will be denoted by $x_{(i)}$, where $x_{(1)}$ is the smallest and $x_{(n)}$ is the largest fragment length. Capital letters, $X$, are used to denoted random variables and small letters, $x$, denote their observed values.

*Tests for a uniform distribution.* The chi-square goodness-of-fit test is computed by dividing the genome into $k$ equally spaced intervals and comparing the observed and expected numbers of sites in each. The statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k}, \tag{1}$$

where $n_i$ is the number of sites in the $i^{th}$ interval and $n$ is the total number of sites. This can be compared to a chi-square distribution with $k-1$ degrees of freedom to determine the significance level. A small p-value would imply that the variation in the counts from different intervals is less than expected for a random distribution of sites. A large p-value would imply that the variation is greater than expected, due to relative excesses or deficiencies of sites within particular intervals.

A variety of distribution-free methods can be used to test the goodness-of-fit to a uniform distribution (13). One advantage of these tests is that they do not depend on an arbitrary assignment of intervals. We have chosen the Kolmogorov-Smirnov statistic which measures the largest difference between the empirical and theoretical distribution functions. The statistic is

$$D_n = \max\left\{ \max_{1 \le i \le n} \left\{ \frac{i}{n} - u_i \right\}, \quad \max_{1 \le i \le n} \left\{ u_i - \frac{i-1}{n} \right\} \right\}. \tag{2}$$

The distribution of $\sqrt{n}D_n$ is tabled (14, table 1.1.2.12) and can be used to determine significance levels. The interpretation of p-values is similar to that for the chi-square test.

One major disadvantage of the Kolmogorov-Smirnov test is that it has poor power properties against general alternatives. A transformation of the data, described by Durbin (15), is based on the ordered fragment lengths and can be used to improve the power of distribution-free tests. The transformed sequence of sites is computed as

$$t_i^* = x_{(1)} + \ldots + x_{(i-1)} + (n+2-i)x_{(i)}; \quad i = 1, \ldots, n. \tag{3}$$

When the observed sites form a Poisson process, so will the transformed sites. The fact that the $t_i^*$ form a Poisson process is a technical matter and derives from properties of the exponential order statistics $x_{(j)}$. The scaled and transformed sites $u_i^* = t_i^*/G$ are used to compute $D_n^*$ as above. Interpretation of the p-values is difficult except to say that extreme values indicate that some interval lengths are more or less frequent than expected.

*Fragment Length Distributions.* Under the Poisson hypothesis, the fragment lengths will be distributed according to an exponential distribution,

$$\Pr(X \le x) = 1 - \exp(-x/\mu), \tag{4}$$

where $\mu$ is the mean fragment length. The corresponding probability density function is the first derivative of the distribution function with respect to $x$. The exponential density function is

$$f(x) = \frac{1}{\mu}\exp(-x/\mu); \quad x \ge 0. \tag{5}$$

The height of the density function at $x$ is proportional to the expected number of fragments with lengths in a small neighborhood around $x$. A maximum likelihood estimate of the parameter $\mu$ is given by the sample mean.

The extreme values from $n$ exponential observations with common mean $\mu$ will be of some interest. A standard probability argument shows the minimum fragment length $X_{(1)}$ will be distributed as exponential with mean $\mu/n$. Thus, the expected size of the smallest fragment will be

$$E(X_{(1)}) = \mu/n. \tag{6}$$

The maximum fragment length, $X_{(n)}$, can be shown to grow at the rate $\ln(n)$. Furthermore, the statistic $X_{(n)}/\mu - \ln(n)$ will have the extreme value distribution

$$\Pr(X_{(n)}/\mu - \ln(n) \le y) = \exp(-\exp(-y)); \quad -\infty < y < \infty. \tag{7}$$

The expected size of the largest fragment is

$$E(X_{(n)}) = \mu(\gamma + \ln(n)), \tag{8}$$

where $\gamma \approx 0.5772$ is Euler's constant. The probability of observing a fragment of size *alpha* or larger can be computed as

$$\Pr(x_{(n)} > \alpha) = 1 - \exp(-\exp(\ln(n) - \alpha/\mu)). \tag{9}$$

Alternative distributions to be considered for the fragment lengths include the gamma and truncated exponential. Both of these distributions will behave as exponentials in the range of large fragments but will have a reduced proportion of small fragments.

The gamma density function can be written as

$$g_{a,\mu}(x) = \frac{x^{a-1}e^{-x/\mu}}{\mu^a\Gamma(a)}; \quad x > 0, \tag{10}$$

where $\Gamma(a)$ is the complete gamma function. The gamma distribution has mean $a\mu$ and variance $a\mu^2$. In the case $a = 1$, it reduces to the exponential. Maximum likelihood estimates of the parameters values can be approximated numerically (16).

The truncated exponential density function can be written as

$$h_{a,\mu}(x) = \frac{1}{\mu}\exp\{-(x-a)/\mu\}; \quad x > a. \tag{11}$$

It has mean $a + \mu$ and variance $\mu^2$ and, for the case $a=0$, reduces to the usual exponential. Maximum likelihood parameter estimates are given by $\hat{a} = x_{(1)}$ and $\mu = \Sigma_{i=1}^{n}(x_i - x_{(1)})/n$. Further details of the distributions and methods presented in this section can be found in (17).

*A Counter Model.* Ideally, we would observe all of the sites and draw our inferences based on complete data. However, as we have described above, real measurements are often imperfect and some sites will go undetected. To model the relationship between the real fragment lengths and the lengths derived from the mapped sites, let us hypothesize the existence of 'locking times', $\{l_i\}$, such that when a site is recorded at position $t_i$, any sites occurring in the interval $(t_i, t_i+l_i)$ will be undetected. The resulting sequence of mapped sites will be a subset of the sequence of true sites such that a true fragment of length less than $l_i$ will be merged into its neighbor.

In the case where the true sites form a Poisson process and the the locking times have a fixed constant value, $l_i = a$, it can be shown that the distribution of observed fragment lengths is truncated exponential. In a long sequence, the expected number of unobserved sites is

$$E(n_{\text{miss}}) = \frac{a}{\mu} n_{\text{obs}}, \qquad (12)$$

where $n_{\text{miss}}$ denotes the number of unobserved sites, $n_{\text{obs}}$ denotes the number of observed sites and the total number of true sites is $n_{\text{tot}} = n_{\text{obs}} + n_{\text{miss}}$. Models with random locking times or locking times proportional to fragment length might appear to be more appropriate here. However, such models are difficult to work with analytically and, as we will describe below, a minor modification of the data makes the fixed locking time model appropriate. A general discussion of counter models can be found in (16).

*Inferences from a Sample* The total numbers of sites in a genome can be inferred from the numbers observed in a sample of sequenced DNA. Assume the true sites follow a Poisson model. If $m$ is the total length of the sample (adjusted for edge effects), we let $r = (G-m)/m$ denote the ratio of unobserved to observed sequence. The expected number of sites in the unobserved portion of the genome will be

$$E(n_{\text{miss}}) = rn_{\text{obs}}. \qquad (13)$$

An approximate standard error of the prediction is given by

$$\text{stderr}(n_{\text{obs}}) = \tfrac{1}{2}\sqrt{4r^2 n_{\text{obs}} + r^2 + 4rn_{\text{obs}}}. \qquad (14)$$

A formal description of this inference can be found in Cox and Hinkley (18, pp. 245).

The same formula for the standard error of prediction can applied to the counter models by letting $r = an_{\text{obs}} / (G - an_{\text{obs}})$ be the ratio of unobserved to observed sequences.

## RESULTS

### Summary Statistics

A summary of the map data is shown in table 2. The numbers of mapped sites, $n$, range from 470 for *Bam*HI to 1567 for *Bgl*I and the corresponding mean fragment lengths $\mu = G/n$ range from 3kb to 10kb. The sample median, $m$, is of some interest as 50% of the fragments are larger/smaller than this value. For exponential fragment lengths, the median will be smaller than the mean. The coefficient of variation $cv = \mu/\sigma$ is expected to be close to 1. Only *Eco*RV departs significantly from 1.

### Distribution of Sites

The chi-square goodness-of-fit test is most useful for detecting local heterogeneity in the frequencies of sites. Values of the chi-square statistic for $k=20$ and 50 are shown in table 3 along with the cumulative chi-square probabilities. The interval sizes corresponding to $k=20$ and 50 are 236kb and 94.4kb respectively. The small p-values obtained for *Bam*HI are very surprising and suggest that the sites are more evenly dispersed than we would expect at random. There are no regions in this size range in the *E. coli* genome which contain either too many

**Table 2:** Summary Statistics for the Physical Map Data

| Enzyme | Site | n | m | $\mu$ | cv |
|---|---|---|---|---|---|
| *Bam*HI | GGATCC | 470 | 6666 | 10043 | 1.047 |
| *Bgl*I | GCC(N5)GGC | 1567 | 2133 | 3012 | 1.093 |
| *Eco*RI | GAATTC | 610 | 5466 | 7738 | 1.066 |
| *Eco*RV | GATATC | 158 | 2267 | 4076 | 0.765 |
| *Hin*dIII | AAGCTT | 517 | 5734 | 9130 | 0.930 |
| *Kpn*I | GGTACC | 497 | 6466 | 9497 | 0.978 |
| *Pst*I | CTGCAG | 846 | 3367 | 5579 | 0.938 |
| *Pvu*II | CAGCTG | 1431 | 2280 | 3298 | 1.064 |

For each enzyme in the physical map, we show the number of recorded sites ($n$), the median fragment length ($m$), the mean fragment length ($\mu$) and the fragment length coefficient of variation ($cv$), defined to be the ratio of the mean to the standard deviation.

**Table 3:** Statistics Testing for a Uniform Distribution of Sites

| Enzyme | $\chi^2_{29}$ | $\chi^2_{49}$ | $\surd nD_n$ | $\surd nD^*_n$ |
|---|---|---|---|---|
| *Bam*HI | 4.213 | 29.26 | 0.594 | 1.291 |
| | (0.00016) | (0.012) | (0.130) | (0.928) |
| *Bgl*I | 14.51 | 44.58 | 0.816 | 5.406 |
| | (0.247) | (0.347) | (0.480) | (>> 0.999) |
| *Eco*RI | 19.97 | 53.77 | 0.539 | 1.463 |
| | (0.630) | (0.703) | (0.068) | (0.972) |
| *Eco*RV | 22.83 | 62.55 | 0.688 | 3.551 |
| (0.755) | (0.908) | (0.272) | (>> 0.999) | |
| *Hin*dIII | 32.36 | 78.45 | 1.180 | 1.319 |
| | (0.972) | (0.995) | (0.877) | (0.938) |
| *Kpn*I | 27.39 | 65.88 | 0.841 | 0.708 |
| | (0.904) | (0.946) | (0.520) | (0.423) |
| *Pst*I | 29.93 | 77.05 | 1.340 | 1.944 |
| | (0.947) | (0.994) | (0.945) | (0.999) |
| *Pvu*II | 24.42 | 47.65 | 1.153 | 1.182 |
| | (0.819) | (0.472) | (0.860) | (0.876) |

The values for the chi-square statistics obtained with the genome divided into 20 segments ($\chi^2_{19}$) and 50 segments ($\chi^2_{49}$) and the Kolmogorov-Smirnov statistics (equation 2) computed on the raw data ($\surd_n Dn$) and the data transformed as in equation 3 ($\surd nD^*_n$) are shown here. In parentheses below each statistic we show the cumulative probability of observing this value under the null hypothesis. Since we are considering two-sided alternatives, very large and very small p-values indicate a departure from the model assumptions.

or too few *Bam*HI sites. The even dispersion is not apparent by visual inspection. Significantly large p-values (> 0.95) obtained for two of the eight enzymes, *Hind*III and *Pst*I, indicate a departure from uniformity in the distribution of these sites. Individual chi-square components were used to identify intervals which made significant contributions to the overall chi-square statistics. The corresponding regions of the map were examined by eye to confirm the excess or deficit of sites indicated. For *Pst*I, a number of high density clusters, spanning 10 to 30kb, were found to be distributed throughout the genome. The most significant of these were located at approximately 290, 1300, 2080 and 2770kb on the Kohara map. For *Hind*III, regions with both excess and deficit of sites were observed. We identified a 190kb region, at approximately 740 to 930kb on the map, which contains only 5 *Hind*III sites and a 120kb region, at approximately 3710 to 3830kb on the map, which contains 26 sites. This latter region includes the primary origin of replication.

Using the untransformed data, none of the Kolmogorov-Smirnov statistics are significant. Although this provides no evidence for departure from uniformity, one should keep in mind the poor power properties of this test. Cumulative plots of the Kolmogorov-Smirnov statistic were used to locate the regions identified above as having significant chi-square components.

When the Kolmogorov-Smirnov test is applied using the transformed data, *Bgl*I and *Eco*RV yield highly significant p-values and the values for *Eco*RI and *Pst*I are also significantly large. The significance of this test statistic indicates that a systematic departure is present in the distribution of the fragment lengths but it gives us little information about the nature of the departure. This suggested that a closer examination of the data would be worthwhile and led us to consider the distribution of fragment lengths in detail, as described in the next section.

In summary, these tests suggest that the cutting patterns produced by six of the eight enzymes depart in some measurably significant way from uniformity. The two enzymes which are not implicated being *Kpn*I and *Pvu*II.

## Fragment Lengths

*Extreme Values* The observed and expected smallest fragment sizes are listed in table 4. In every case, the observed smallest fragment size is significantly larger than its expectation. It should be noted that the observed values are sensitive to errors from a number of sources. The recorded sizes are below the resolution limits of a typical gel and also reflect periodicities introduced at the data extraction stage. Experimental errors on these measurments are probably greater than 100%. For the three enzymes with the smallest mean fragment sizes ( *Bgl*I, *Pvu*II and *Eco*RV) the expected minimum fragment size is actually smaller than the recognition pattern. As none of these patterns can self-overlap, such small fragments would never occur. These observations point out that the Poisson approximation is least reliable in the range of small fragment sizes and provide strong evidence for a departure from the Poisson assumptions in the map data. It should be added, for reasons to be discussed, that we believe that the true smallest fragments are in fact smaller than those observed in the data.

The maximum fragment length is of some practical interest in the construction of libraries from partial digests with a single enzyme (19). The observed and expected maximum fragment lengths shown in table 4 are in close agreement except *Eco*RV, for which the largest fragment is more than double its expected value. The probability of observing a fragment of this size or larger is, by equation 9, equal to 0.00018. Even after correction for the multiple testing implicit in choosing the most extreme of eight observations, the attained significance level is less than 0.0015. This fragment is indeed larger than we would expect. It should be noted that all of the 10 largest *Eco*RV fragments span regions designated as uncertain on the Kohara map. Thus, they probably represent failure of the partial digest reactions and not truly large fragments. The 76kb *Kpn*I fragment, the largest observed over all eight enzymes, is well within the range of expectations ($p = 0.150$).

*Estimated Fragment Length Distributions* Under the Poisson hypothesis, the fragment lengths should follow an exponential distribution. We have already seen evidence that the sites represented in the map depart from this model. In figure 1 we show, for each enzyme, plots of the fragment length densities estimated by a non-parametric Gaussian kernal method (20, 21 pp. 295) with the best fitting exponential density superimposed. The height of these curves is roughly proportional to the numbers of fragments in a small neighborhood around the given length. A consistent pattern of departure is seen as a deficiency of small fragments in the 0 to 2kb range. Unfortunately, the non-parametric density estimate is unreliable near the boundary at zero. The exponential fit appears to be good in the upper tail, corresponding to longer fragment lengths, with the exception of *Eco*RV for which too many very large fragments are observed.

*Alternative Fragment Length Models* These observations suggest that we should consider alternatives to the exponential model for fragment lengths. Two good choices are the gamma and truncated

**Table 4:** Sizes of Extreme Fragments

| Enzyme | Smallest | | Largest | |
|---|---|---|---|---|
| | Observed | Expected | Observed | Expected |
| *Bam*HI | 267 | 21.4 | 65746 | 67589 |
| *Bgl*I | 133 | 1.9 | 21106 | 23899 |
| *Eco*RI | 134 | 12.7 | 41067 | 54088 |
| *Eco*RV | 133 | 3.6 | 63866* | 31105 |
| *Hind*III | 400 | 17.7 | 62800 | 62311 |
| *Kpn*I | 147 | 19.1 | 76253 | 64449 |
| *Pst*I | 54 | 6.6 | 44173 | 40830 |
| *Pvu*II | 147 | 2.3 | 29840 | 31151 |

The observed smallest and largest fragment sizes were extracted from the Kohara map data and expected values were computed as described in the text. It should be noted that the smallest fragment sizes are sensitive to periodicity introduced at the data extraction stage.

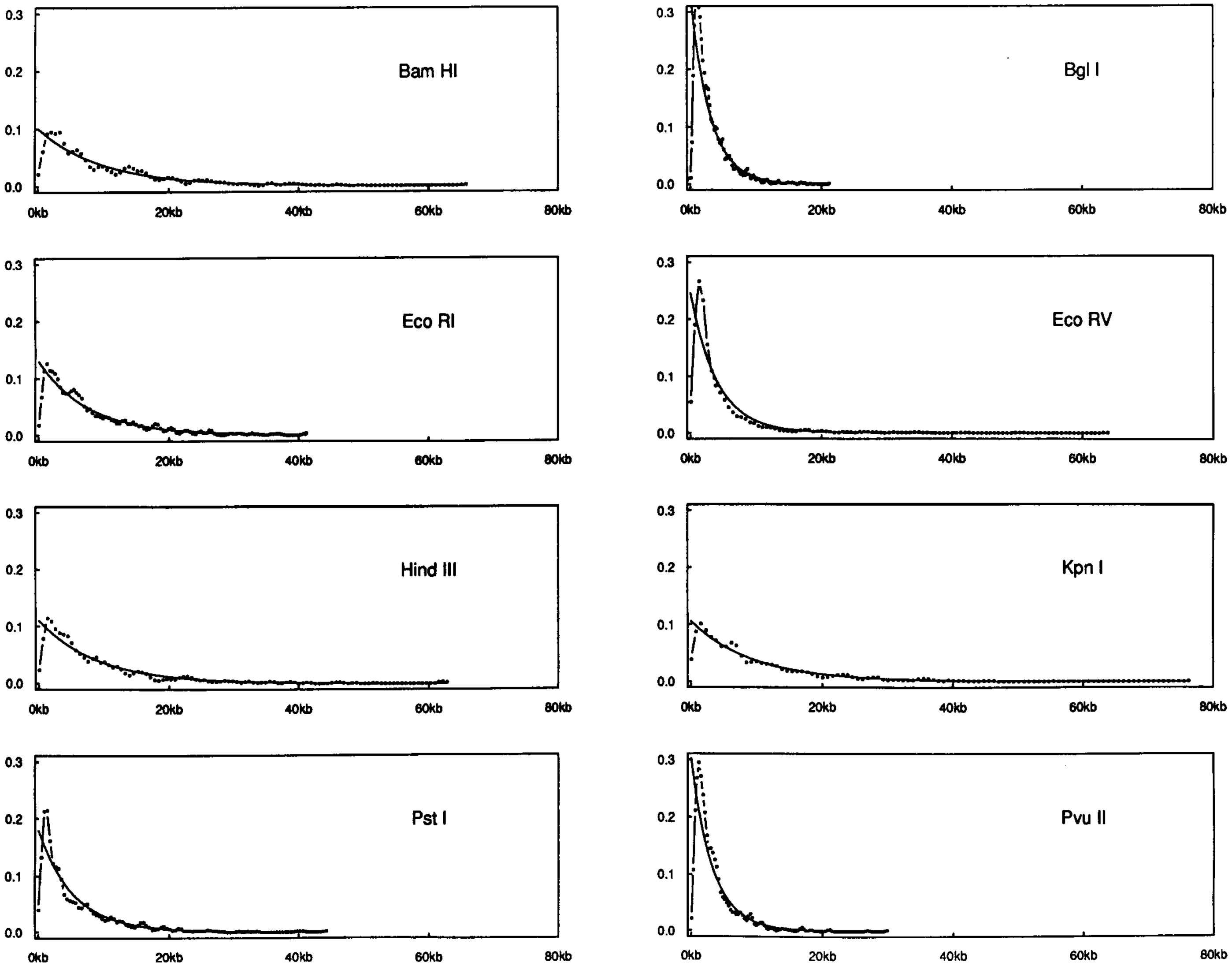

**Figure 1:** Estimated Densities of Fragment Lengths. A Gaussian kernal smoothing algorithm was used to estimate the density of fragment lengths at 100 equally spaced points from zero to the maximum fragment length for each of the eight enzymes. The results are plotted as a connected series of points. The height of the curve is proportional to the numbers of fragments with lengths falling in a weighted neighborhood of the plotted point. Exponential density functions with mean parameter, $\mu$, estimated by maximum likelihood are shown superimposed as smooth curves. The horizontal scale for plotting is fragment length and the vertical scale is chosen to make the integrated area under the curves unity.

exponential models. Both contain the exponential as a special case, require only one additional parameter and are analytically tractable. They are consistent with the observed data in that the upper tails of their density functions behave as an exponential while the lower tails are lighter. The suppression of small fragments under a gamma model is gradual while for the truncated exponential it is abrupt. Both alternatives are suggested by the paucity of small fragments.

Likelihood ratio tests reject the exponential model in favor of both alternatives for all eight enzymes. The enzymes *BglI* and *PstI* have higher likelihood values for the gamma model and the remainder favor the truncated exponential model. We conclude that the best description of the fragment length distribution is not exponential. However, there are some problems with the alternative models as well. The truncated exponential relies heavily on the minimum fragment length which we know to be poorly determined. The loss of small fragments does not appear to be absolute as this model would imply. The detection of small fragments is likely to be highly variable due to variations in gel resolution and other factors involved in mapping experiments. Examination of histograms for fragments less than 2kb in size (not shown) reveals that the truncation is not abrupt. For the gamma model we have no good physical justification. In the next section, we will describe a modification of the data which makes the truncated exponential model more suitable and allows us to predict the numbers of sites missing from the map data.

## Finding the Missing Sites

Because of variations in individual gels, there is no reason to believe that an absolute threshold exists for the resolution of closely spaced sites. Modeling of counter processes with non-constant locking times is analytically complex and would require additional assumptions about the structure of errors in the map data. To circumvent these problems, we will make the following modification of the data. Assume that fragments larger than some given size, $a$, are always resolved as distinct bands on a gel. Observed fragments with sizes smaller than this limit will be merged into their neighboring fragments by adding lengths, as

**Table 5.** Predictions of the Actual Numbers of Sites in *E. coli.*

| Enzyme | Kohara | 1kb Counter | 2kb Counter | Sample |
|---|---|---|---|---|
| *Bam*HI | 470 | 494 | 493 | 442 |
| | | (7.2) | (10.2) | (97.1) |
| *Bgl*I | 1567 | 1804 | 1781 | 1946 |
| | | (26.3) | (36.7) | (203.0) |
| *Eco*RI | 610 | 644 | 646 | 773 |
| | | (9.4) | (13.3) | (128.2) |
| *Eco*RV | 1158 | 1267 | 1163 | 2077 |
| | | (18.5) | (23.9) | (209.6) |
| *Hin*dIII | 517 | 541 | 529 | 840 |
| | | (7.9) | (10.9) | (133.5) |
| *Kpn*I | 497 | 507 | 504 | 552 |
| | | (7.4) | (10.4) | (108.5) |
| *Pst*I | 846 | 872 | 836 | 1384 |
| | | (12.7) | (17.2) | (169.0) |
| *Pvu*II | 1431 | 1663 | 1610 | 2364 |
| | | (24.2) | (33.1) | (223.6) |

The column labeled Kohara shows the numbers of sites observed in the physical map data. Predictions of the total numbers of sites based on counter models with minimum fragment sizes of 1kb and 2kb are shown in the second and third columns. The final column shows the predicted numbers of sites based on a sample of sequenced *E. coli* DNA. Values in parentheses are standard errors for these predictions.

if the site had actually not been detected. In this way, we coerce the data to fit a counter model with fixed and known locking times. The amount of information lost by this procedure is small but increases as $a$ is increased. The result is an increase in the variance of the estimated $n_{\mathrm{miss}}$. As $a$ is decreased below the worst case resolution of a gel, our model assumptions begin to fail and a downward bias is introduced into the estimate. Based on plots of $n_{\mathrm{miss}}$ versus $a$, we have chosen $a=1$kb and $a=2$kb as representative values which acheive low variance and low bias for most of the eight enzymes.

Counter model estimates of the total numbers of sites are shown in table 5. The largest increases over the observed numbers of sites are seen for enzymes with the highest frequencies of cutting. The proportions of missing to total sites range from less than 5% for low frequency cutters to over 15% for the high frequency cutters. The two estimates are in generally good agreement. The notable exception is *Eco*RV, for which the 1kb model predicts 109 additional sites and the 2kb predicts only 5. This is explained by the increased influence in the estimates of the very large fragments under the 2kb model.

Independent estimates of the total numbers of sites were obtained by counting the sites in sample of sequenced *E. coli* DNA representing 4.5% of the total genome. Predictions and their standard errors are shown in table 5. Again, we see that the largest numbers of missed sites are predicted for the frequent cutters. The sample estimates, with the exception of *Bam*HI, are higher than the the counter estimates but most are close to or within the range of 2 standard errors. The exceptions are *Eco*RV, *Pst*I and *Pvu*II. In the case of *Eco*RV, we suspect this is due to the failure of the counter model to adequately predict the missing sites. Heterogeneity in the sample, i.e. excess variance in the numbers of sites per fragment, was observed for *Pst*I ($\chi^2_{29} = 72.4, p >> .999$) and to a lesser extent for *Kpn*I and *Pvu*II. The estimates of $n_{\mathrm{tot}}$, for these three enzymes, are still unbiased but the reported standard errors may be too small.

### Sequence to Map Alignments

Three segments of sequenced E.coli DNA (ECOGLTA, ECORPLRPO and ECOUNCC) were converted to eight enzyme restriction maps and aligned to the Kohara map using a modification of the algorithm of Waterman et al. (22) suitable for fitting map segments into a longer map. The alignments were then modifed by hand to allow for ambiguous orderings of closely spaced sites and multiple to one matchings. Multiple matching was introduced in another paper (23) and could easily be incorporated into these algorithms. Sequence to map alignments are likely to be very useful tools for the assembly of genomic sequence data and refinements of the algorithm which allow for these common mapping errors are being considered. The paper of Rudd et al. (2) presents a distinct approach to the problem of locating sequence segments on a map. The alignments shown in figure 2 confirm that multiple closely spaced sites are frequently represented as single sites on the Kohara map data. Of the 96 mapped sites found in these alignments, 8 appear to be double and 3 appear to be triple sites in the sequence data. This result confirms the conclusion drawn from the counter model analysis of the map data. Approximately 10% of mapped sites represent multiple closely spaced sites in the DNA sequence.

## CONCLUSIONS

In this paper, we have examined the distribution of restriction sites in the *E. coli* genome using data recorded in an extensive physical map. Statistical methods of analysis are based a top-down modeling approach which requires no knowledge of statistical properties of the base sequence. Our main conclusion is that, after taking into account the missing sites, the Poisson model provides a good approximation to the actual distribution of these eight restriction sites in *E. coli*. This result is perhaps not surprising, as the Poisson model is often implicitly assumed for the distribution of restriction enzyme sites. We feel however that the Poisson model will fail in many cases, especially for higher organisms, and the methods described here could provide a starting point for more detailed analyses.

The distribution of sites for most of the eight enzymes studied appears to be fairly homogeneous throughout the *E. coli* genome. However, the *Bam*HI sites are too evenly dispersed. The presence of the the Dam methylation site within the *Bam*HI recognition

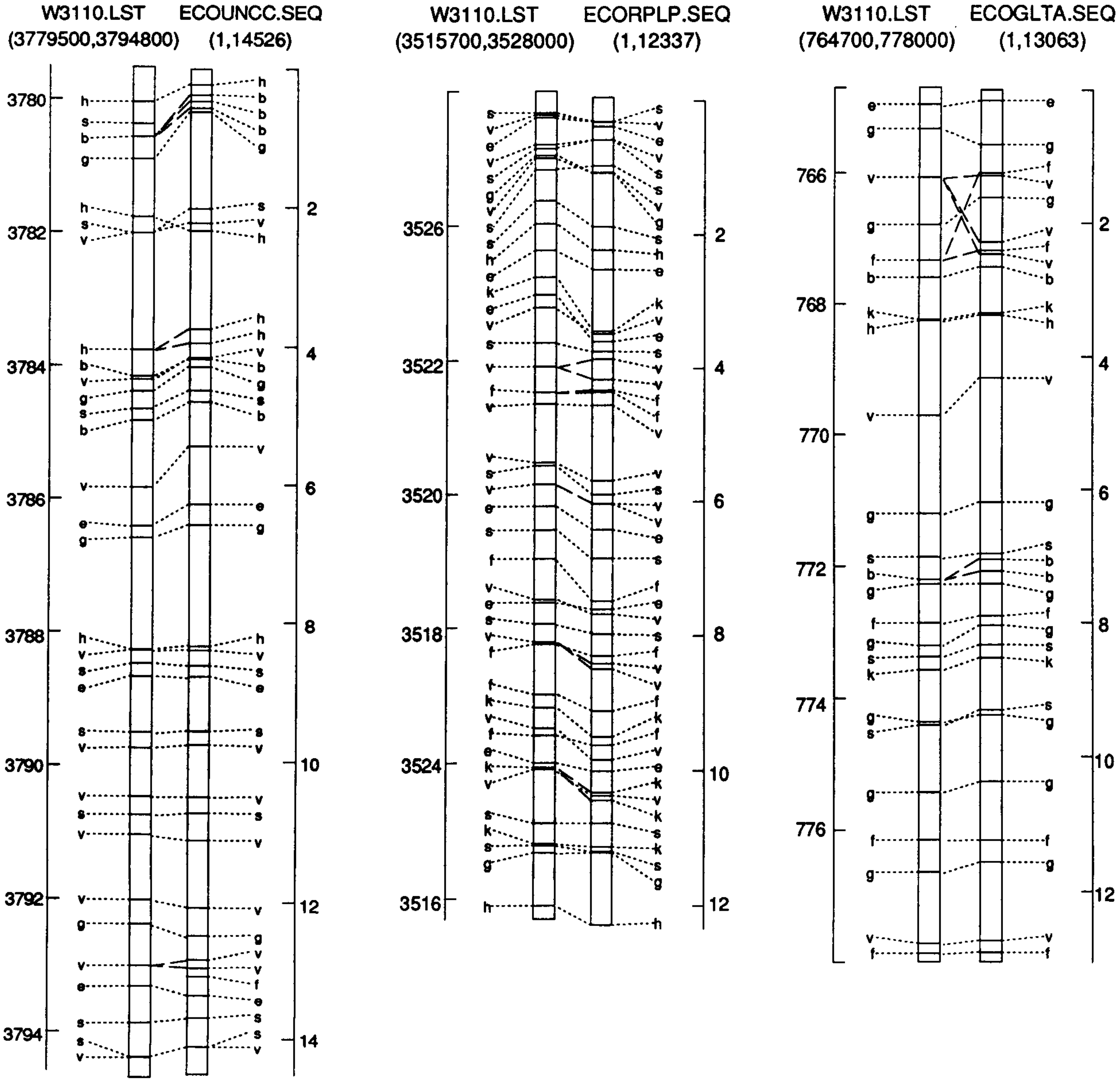

**Figure 2:** Sequence to Map Alignments. Restriction maps predicted from DNA sequence taken from GenBank are aligned with the corresponding region of the W3110 restriction map. Dotted lines show one-to-one correspondence between sites and dashed lines show the correspondence between multiple sites of the same enzyme which had been mapped as a single site in the W3110 map. Coordinate numbers in kb for the W3110 map are indicated. Coordinate numbers for maps derived from sequence are in kb with the first base of the sequence defined as .001. Abbreviations used are: b = *BamI*, g = *BglI*, e = *Eco*RI, f = *Eco*RV, h = *Hin*dIII, k = *KpnI*, s = *PstI* and v = *PvuII*.

pattern suggests that a functional constraint may be involved, possibly related to the need for excision repair functions at fairly even intervals throughout the genome. Other known functions of Dam methylation include a role in the timing between DNA replication initiation events (24). Heterogeneity in the distribution of *PstI* sites, seen in both the physical map and sample sequences, appears to be due to presence of several small clusters (<50kb) of sites. For *Hin*dIII, the relative excesses and deficiencies of sites cover larger regions (>100kb) of the genome. A detailed study of local variations in restriction site frequencies may yield further insights, but for most purposes homogeneity would seem to be a reasonable assumption.

The main departure from a uniform distribution of the sites observed in the map data is due a deficiency of small fragments. We conclude, based on knowledge of the mapping procedures and analysis of the map data that closely spaced sites are actually present in the sequence but are undetected in the map. The quality of the map is excellent in general but one should keep in mind that 5 to 15% of the mapped sites represent multiple sites in the DNA sequence.

Results obtained using the counter models appear to be generally reliable, with the exception of *Eco*RV. We conclude, based on the sampling estimates and the uncertainty indicated in much of the Kohara map for this enzyme, that *Eco*RV sites are distributed at a high frequency throughout the genome, with perhaps more than 2000 sites. The failure of the counter model to properly predict this number is due the presence of large apparent fragments in the map data. These in turn are due the failure of partial digest reaction for many clones and are not explicitly accounted for by the model.

## ACKNOWLEDGEMENTS

The authors wish to thank Kenn Rudd for sharing data and results prior to their publication and also the anonymous referees for comments which helped improve the qualtiy of the presentation.

This research was supported by grants GM36230 and GM35682 from the National Institutes of Health and grant DMS-8720208 from the National Science Foundation

## REFERENCES

1. Kohara Y., Akiyama K., Isono K. (1987) Cell **50**:495–508.
2. Rudd K.E., Miller W., Ostell J., Benson D.A. (1989)
3. Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., Rodier F. (1985) Science **228**:953–958.
4. Elton R.A. (1974) J. Theor. Biol. **45**:533–553.
5. Messer W., Noyer-Weidner M. (1988) Cell **54**:735–737.
6. McClelland M., Jones R., Patel Y., Nelson M. (1987) Nucl. Acids Res. **15**:5985–6005.
7. Biggins J.D. and Cannings C. (1987) Adv. Appl. Prob. **19**:521–545.
8. Breen S., Waterman M.S., Zhang N., (1985) J. Appl. Prob. **22**:228–234.
9. Bishop T., Williamson J.A., Skolnick M.H. (1983) Am. J. Hum. Genet. **35**:795–815.
10. Waterman M.S. (1983) Nucl. Acids Res. **11**:8951–8956.
11. Arratia R., Goldstein L., Gordon L. (1989) Ann. Prob.
12. Smith H.O. and Birnsteil M.L. (1976) Nucl. Acids Res. **3**:2387–2400.
13. Lewis P.A.W. (1965) **52**:67–77.
14. Bronshtein I.N. and Semendyayev K.A. (1985) Handbook of Mathematics. Van Nostrand Reinhold, New York.
15. Durbin J. (1961) Biometrika **48**:41–55.
16. Bowman K.O. and Shenton L.R. (1987) Properties of Estimators for the Gamma Distribution. Marcel Dekker, New York.
17. Karlin S. and Taylor H.M. (1975) A First Course in Stochastic Processes. Academic Press, New York.
18. Cox D.R. and Hinkley D.V. (1974) Theoretical Statistics. Chapman and Hall, New York.
19. Tang B. and Waterman M.S. (1989) Bull. Math. Biol.
20. Wegman E.J. (1972) **14**:533–546.
21. Becker R.A. and Chambers J.M. (1984) S -- An Interactive Environment for Data Analysis. Wadsworth, Belmont, CA.
22. Waterman M.S., Smith T.F., Katcher H.L. (1984) Nucl. Acids Res. **12**:237–242.
23. Waterman M.S. and Raymond R.Jr. (1987) **19**:109–127.
24. Bakker A. and Smith D.W. (1989) J. Bacteriol. **171**:5738–5742.