© 1991 Oxford University Press   Nucleic Acids Research, Vol. 19, No. 15 4241–4246

Assessment of inhomogeneities in an *E.coli* physical map

Samuel Karlin and Catherine Macken
Department of Mathematics, Stanford University, Stanford, CA 94305, USA

Received March 19, 1991; Revised and Accepted June 14, 1991

ABSTRACT

A statistical method based on *r*-fragments, sums of distances between (*r* + 1) consecutive restriction enzyme sites, is introduced for detecting nonrandomness in the distribution of markers in sequence data. The technique is applicable whenever large numbers of markers are available and will detect clumping, excessive dispersion or too much evenness of spacing of the markers. It is particularly adapted to varying the scale on which inhomogeneities can be detected, from nearest neighbor interactions to more distant interactions. The *r*-fragment procedure is applied primarily to the Kohara *et al.* (1) physical map of *E.coli*. Other applications to DAM methylation sites in *E.coli* and NotI sites in human chromosome 21 are presented. Restriction sites for the eight enzymes used in (1) appear to be randomly distributed, although at widely differing densities. These conclusions are substantially in agreement with the analysis of Churchill *et al.* (3). Extreme variability in the density of the eight restriction enzyme sites cannot be explained by variability in mono-, di- or trinucleotide frequencies.

INTRODUCTION

A physical restriction site map of the *E.coli* (K12 strain) genome has been presented recently (1). The Kohara map of *E.coli* was generated via partial digestion of DNA using eight different six-cutters. The map has been digitized by K.Rudd and colleagues of the National Institutes of Health, Washington DC, who kindly supplied us with the data (2). Table 1 contains some summaries of these data. What insights on the nature of local and global heterogeneities of the *E.coli* genome emerge from the Kohara map? Churchill *et al.* (3) previously extensively investigated the distribution of restriction sites in an alternate digitization of the Kohara map. We present here a new statistical technique with broad applicability by means of which we shall study the *E.coli* map.

In probing for insights on the organization of the *E.coli* or other genome, the general problem arises of how to characterize anomalies in the spacings of markers in a long sequence of nucleotides or amino acids. In particular, how does one assess unusual features such as excessive clumping (too many neighboring short spacings), overdispersion (too many long gaps between markers), or too much regularity (too few short spacings and/or too few long gaps)? Specific questions are: does a dense clump of markers indicate a significant departure from the random model and, is a large gap indicative of a region with significant underrepresentation of markers? Churchill *et al.* (3) earlier addressed these questions.

Our initial examination of the *E.coli* physical map led to an intriguing observation: On the one hand, the arrangement in the genome of restriction sites of each of the eight restriction enzymes appears homogeneous (consistent with a random distribution). Specifically, apart from small sizes, the collection of fragment lengths for each enzyme type can be well described by the distances between sites sampled randomly from a uniform distribution over the genome (see below for details). This finding is substantially equivalent to conclusions in (3). On the other hand, the counts of the different restriction sites vary widely from 470 up to 1572 (see Table 1). Explanations of the diversity in counts in terms of mono-, di- or trinucleotide genome content were unsatisfactory (see below). Other factors must therefore be responsible for the large variance in counts.

Questions about spacings of a marker sequence and a general interest in sequence heterogeneity led us to a consideration of the lengths of groups of *r* (e.g., *r* = 2,3,5,10) consecutive fragments (henceforth called ***r*-fragments**), where a fragment length is the distance (measured in bases for DNA, measured in residues for proteins) between two adjacent marker sites. In particular, we focus on the *k* longest and *k* shortest lengths (where, for example, *k* = 3) of *r*-fragments, as appropriate statistics for detecting cases of excessive clumping, excessive gaps, or excessive regularity in the spacings of the marker. The case *r* = 1, *k* = 1, corresponds to analyses of fragment lengths in (3). The use of sums of *r* consecutive fragment lengths, rather than single (*r* = 1) fragment lengths, provides greater sensitivity for detecting unusual spacings in the marker array. The *r*-fragment statistics are also better able to tolerate measurement errors and reduce effects of statistical fluctuations than single fragment lengths. Our analyses using *r*-fragments led to a number of interesting findings including overdispersion of EcoRV sites in some regions of the genome, a large clump of PstI sites, and apparent regular spacings of BamHI sites. The findings corroborate some of those reported earlier in Churchill *et al.* (3). However, our testing is at a more stringent level than that of (3) to allow for multiple tests, and thus we do not report as many significant observations as this earlier work. We believe that our methods have potential relevance to future restriction mapping efforts.

Table 1. Summary of *E.coli* physical map data

Enzyme Name	Recognition Sequence	Number of Sites	Extreme fragment sizes (bp) Minimum	Maximum
BamHI	GGATCC	470	400	65,700
BglI	*GCCNNNNNGGC	1572	100	21,300
EcoRI	GAATTC	613	300	40,800
EcoRV	GATATC	1159	100	64,000
HindIII	AAGCTT	518	400	62,800
KpnI	GGTACC	497	300	76,600
PstI	CTGCAG	848	100	44,200
PvuII	CAGCTG	1435	100	29,200

*N denotes: Any nucleotide may be inserted in this position. Number of sites and extreme sizes for single fragments generated by eight restriction enzymes as recorded in a digitized version (2) of the physical map of *E.coli* constructed by Kohara *et al.* (1).

METHODS

Counts of markers

Mono-, di- and trinucleotide binomial model predictions: For a nucleotide sequence generated by *independently* selecting successive letters according to a set of prescribed probabilities per letter type, the likelihood of observing a particular 6-word (hexanucleotide) at any specified location is the product of the frequencies of the component letters.

For a dinucleotide (*nearest neighbor dependence*) model, the probability of observing a particular 6-word, say w = CAGCTG is taken to be

$$fw = \frac{fCAfAGfGCfCTfTG}{fAfGfCfT}$$

For a trinucleotide (*two-neighbor dependence*) model, the probability of w would be assessed by

$$fw = \frac{fCAGfAGCfGCTfCTG}{fAGfGCfCT}$$

In all of these binomial models, the expected count of the word w is $N f_w$ where N is the length of the sequence. These binomial models have been used widely (see, for example, (4−7)).

Markov chain estimates: For a sequence of letters generated as a first-order Markov chain (letter X follows letter Y with probability $f_{X|Y}$), we ascertain the statistics of the distance (number of letters) between consecutive occurrences of the marker (i.e., the first passage distance). The calculation of the mean μ and variance σ^2 of the first passage distance is arduous but standard (see, for example, chaps. 3 and 5 of (8)). The expected count of markers in the long sequence is then about L/μ (L is the length of the sequence) with variance $L\sigma^2\mu^3$ (see loc. cit., p. 208). An analogous second-order Markov chain model can be constructed using trinucleotide frequencies.

For the *E.coli* physical map, all the above models of randomness led to discordant results compared to the actual counts (see below and Figure 1).

Distribution of lengths of fragments and *r*-fragments

Consider a sequence of N letters. Suppose a marker with n occurrences is located throughout the sequence. Occurrences of markers induce $n + 1$ **spacings** (fragment lengths), $U_0, U_1, \ldots, U_n$, where U_i is the distance (number of letters) between the i^{th} and $(i + 1)^{st}$ marker ($i = 1, 2, \ldots, n - 1$); U_0 is the distance until the first marker; U_n is the distance between the last marker and the end of the sequence. We scale distances so that one letter has a length of $1/N$ units. Now we can regard the successive spacings $\{U_i\}_{i=0}^{n}$ as a partition of the unit interval. We calculate all r-fragment lengths ($r = 1, 2, 3, \ldots$)

$$R_i = \sum_{j=i}^{i+r-1} U_j, \qquad i=1,2, \ldots, n-r+1 \qquad [1]$$

Note that in the case of circular genomes, such as *E.coli*, n markers induce n spacings. Our methods are essentially unaffected by the discrepancy, since we shall be considering only the case of large n.

To study organization of the sequence, we compare the distribution of $\{U_i\}$ and $\{R_i\}$ under a postulated model for randomly located markers with the observed distribution of fragment and r-fragment lengths. The extreme-valued fragments (largest and smallest) are of particular use in detecting inhomogeneity. We consider the following:

$$M_k{}^{(r)} = k^{\text{th}} \text{ largest } \{R_i\}, \qquad [2]$$

$$m_k{}^{(r)} = k^{\text{th}} \text{ smallest } \{R_i\}, \qquad [3]$$

for, say, $k = 1, 2, 3$.

Our r-fragment test proceeds as follows: To detect the clumping of markers, we check whether the minimum r-fragment length is especially small for the postulated random distribution of markers. Similarly, to decide if any grouping of markers are excessively separated, we check whether the maximum r-fragment length is especially large. Conversely, when the minimum r-fragment length is especially large and/or the maximum r-fragment length is especially small, then the spacings of the marker would be judged overly regular. The case for $r = 1$ is classical, and the statistics when $r = 1$, $k = 1$ are equivalent to the maximum and minimum fragment lengths considered in (3).

To assess overdispersion, we use the theoretical probability that the k^{th}-largest r-fragment length, $M_k{}^{(r)}$ would be as large or larger than those observed if markers were in fact distributed randomly (sampled uniformly over the long sequence):

$$Pr\{\mathrm{M}_{\mathrm{k}}^{(\mathrm{r})} > \frac{1}{n}[\ln n+(r-1)\ln(\ln n)+x]\}$$

$$\approx 1-\exp(-\mu)\left\{\sum_{i=0}^{k-1} \frac{\mu_i}{i!}\right\}, \ \mu=\frac{e^{-x}}{(r-1)!} \qquad [4]$$

With x chosen so that the right side of Equation [4] is equal to 0.01, we declare the observed $M_k{}^{(r)}$ 'too large' if it exceeds $\frac{1}{n}[\ln n + (r - 1) \ln(\ln n) + x]$. The conservative (1%) level of significance is advised because of the large number of tests being conducted.

To assess clumping, we use the theoretical probability that the k^{th}-smallest r-fragment length, $m_k{}^{(r)}$ would be as small or smaller than those observed if markers were in fact located randomly:

$$Pr\left\{m_k^{(r)} < \frac{x}{n^{1+1/r}}\right\} \approx 1-\exp(-\lambda)\left\{\sum_{i=0}^{k-1} \frac{\lambda^{\mathrm{i}}}{i!}\right\}, \ \lambda=\frac{x^r}{r!} \qquad [5]$$

With x chosen so that the right side of Equation [5] is equal to 0.01, we declare the observed $m_k^{(r)}$ 'too small' if it is less than $\chi/n^{1+1/r}$.

To detect too much regularity, we use the theoretical probabilities that $M_k^{(r)}$ are especially small, and $m_k^{(r)}$ are especially large, calculated from Equations [4] and [5].

r-fragments ($r > 1$) are more sensitive detectors of nonrandomness than mere fragment lengths ($r = 1$) since by summing lengths, the magnitude of measurement errors relative to natural fluctuations in the sums is reduced. Furthermore, varying r allows us to modulate the scale at which we may detect inhomogeneities. The distributional formulas for extreme-valued *r*-fragments were first derived in (9) and (10) for $k = 1$ with (11) and (12) giving the general case. In these technical references the term *r*-scan is used in place of *r*-fragment.

RESULTS

E.coli physical map data

The digitized Kohara physical map data contains errors of two kinds: measurement error, and experimental error. Part of the measurement error arises from the process of digitizing an enlarged copy of the Kohara physical map where locations of restriction sites were recorded to the nearest 100 bp. Thus, as a result of digitizing, no two sites of the same type are separated by a distance of less than 100 bp, while sites of different types might be separated by 0 bp. Insight into the true distances separating sites is afforded by sequence data. About 1.43×10^6 bp of sequence data were made available to us with the kind permission of K.Rudd and are now accessible in the public domain (13). Data were cleaned by K.Rudd *et al.* (13) to remove redundancies in the original (GenBank) data source. For example, in the 80 sequenced fragments or contigs longer than 5000 bp, about 3% of the distances between consecutive PstI sites are less than 100 bp; the comparable figure for consecutive EcoRV sites is approximately 6%. The coarseness of the digitized data affects the ability to detect clumping of sites (see below).

Experimental error takes several forms and is discussed in detail by Churchill *et al.* (3). Errors of particular concern to our analysis are: imprecise measurement of fragment lengths; failure to distinguish two closely-spaced sites of the same type; and possible inversion of the order of two closely-spaced sites of different types. In addition, some stretches (approximately 10%) of the map were missing EcoRV sites, due to difficulties in reading autoradiograms. In (2) and (3), alignment of sequence data with the physical map provides some checks on the quality of the physical map.

Counts of restriction sites in *E.coli*

Sites are underrepresented: We scanned about 1.43×10^6 bp of *E.coli* DNA sequence data (13) for the presence of all the Kohara restriction sites except BglI. The counts obtained were scaled by 4.7196/1.4316 with the outcome presented in Table 2. The three most abundant sites, EcoRV, PstI and PvuII, were detected in the physical map at a level 70% or less than that predicted from the scaled sequence data. This underrepresentation probably reflects the inability to distinguish short fragments of similar lengths in the gel. Churchill *et al.* (3) used an elegant counter model to predict the number of missing sites. In all cases except one (BamHI), their predicted total number of sites was less than that estimated from a sample of sequence data appropriately scaled. Further evidence that some clumping of sites was undetected by the physical map is provided by counts of pairs of sites, and mean distances between consecutive sites. For all types of sites, the counts of consecutive pairs of like type are significantly low, as measured by a log odds ratio test (data and analysis not shown). This anomaly in ordered restriction site pair distances is not apparent in the available *E.coli* sequence data.

Table 2. Comparison of observed and predicted site frequencies

Enzyme Name	Observed	Expected			
		Sequence	Binomial Model		
			Mono	Di	Tri
BamHI	470	466	1220	912	1106
EcoRI	613	833	1080	1497	1361
EcoRV	1159	2160	1080	803	1401
HindIII	518	675	1080	1386	1318
KpnI	497	560	1220	583	1279
PstI	848	1238	1220	1304	2562
PvuII	1435	2025	1220	1304	3337

For each enzyme in the physical map, except BglI, the observed frequency of restriction sites in the map is compared with that predicted on the basis of scaled sequence data or a binomial model using mono-, di-, or trinucleotide frequencies extracted from available sequence data (13).

Poor predictions of counts of sites: We estimated mono-, di- and trinucleotide probabilities from 1.43×10^6 bp of sequenced *E.coli* DNA (2). The various binomial models predicted counts of restriction sites in discord with observed counts and each other (see Table 2 and Figure 1). Predictions from Markov models were essentially identical to those of the binomial models. In particular, none of the models predicted counts of the different markers with the same rank ordering as that of the physical map. While the binomial model using trinucleotide frequencies is a poor quantitative predictor for the seven Kohara 6-cutter counts, the predicted counts have the same rank order as the observed counts except for the inversion of EcoRV and PstI. Dinucleotide frequencies in the binomial model yielded a reasonable estimate of the count of KpnI sites, and matched the scaled sequence count of PstI sites. Table 2 underscores the effect of neighbor dependencies on the predicted number of sites. For example, recognition sequences for KpnI and PstI contain the same number of each of the nucleotides, although in different orders; based on dinucleotide frequencies, they produced very different expected counts (583 and 1304, respectively).

Distribution of lengths of fragments and *r*-fragments in *E.coli*

Sequence data shows homogeneous locations of markers: The DNA sequence data afforded a rough examination of the density of sites around the genome (and thereby, the spacings of sites). We compared the density of sites per unit length in each of the long (≥ 5000 bp) fragments or contigs contained in the sequence data (13). Using a chi-squared test of homogeneity, we found no evidence of atypical regions of the genome (data and analysis not shown).

Fragment lengths approximately exponentially distributed: The quality of the data is such that test results were difficult to interpret with confidence. For each type of restriction site, histograms of the fragment lengths revealed roughly exponential tails for lengths greater than about 2000 bp. However, compared with the exponential distribution, we noted an underrepresentation of short fragments, almost certainly due to the coarseness of the physical and digitized maps. Using these histograms, and chi-squared goodness-of-fit tests, we conclude that fragment lengths have a distribution similar to the exponential distribution, thus suggesting homogeneous spacings of sites. Our results are in agreement with Churchill *et al.* (3).

Over-dispersion of EcoRV fragments: Because of sample size, we were restricted in testing for overdispersion of individual restriction sites to *r*-scans with $r \leq 2$. The first three maxima of 2-scans of EcoRV recognition sequences were significantly large at the 1% level (see Table 3). These three significant maxima all fell at positions in the genome where EcoRV sites are missing. Our tests found no signs of overdispersion of any other recognition sequence.

Large clump of PstI sites: To test for clumping, we could use *r* as large as 10 without compromising the applicability of the asymptotic test. For $r = 10$, *r*-fragments detected significant clumping of PstI sites that was not found with $r = 5$ (see Table 3). The first three minima were all significant and their positions overlapped. These minima localize a clump of 13 PstI sites beginning at map position 2074.8 kb and spanning 13 kb. Four of these PstI sites had no intervening site of any of the other seven types. A further two alternated with KpnI sites. Our conclusions differ somewhat from those of Churchill *et al.* (3) who claimed a number of clumps of PstI sites. However, their testing was at a lower level of stringency than ours. The cluster located by our *r*-fragment procedure probably corresponds to one located at 2770 kb in (3).

Suggestion of regularity in BamHI sites: When we tested for especially regular spacing, we found that the majority of the minima were 'too large'. We were probably measuring the effects of the coarseness of the data. We might expect that the bias due to the coarseness of the data would diminish for large *r*. Since we found significantly large minima for the least frequent cutter BamHI even at $r = 10$, we proffer that BamHI sites are considerably more evenly spaced than expected if its sites were distributed randomly throughout the genome. Churchill *et al.* (3) earlier noted the possibility of regularity in the spacing of BamHI sites.

r-fragments are sensitive detectors of non-randomness: Table 4 gives threshold values for representative numbers of sites in the *E.coli* genome for the detection of clumping at the 1% and 5% level of significance. For example, if we were concerned with a restriction enzyme which had 1000 recognition sites in the genome, the minimum distance between pairs of sites ($r = 1$) in order to declare 'significant' clumping is much less than 1 bp! This threshold is clearly lower than the limits of detection imposed by any map. From this perspective, the usefulness of *r*-fragments is apparent.

Other applications of the *r*-fragment procedure

Our *r*-fragment procedure is broadly applicable. Motivated by discussions during the preparation of this paper, we carried out two further tests of randomness of markers with the following results.

Cluster of GATC sites in the ori-C region of E.coli? A group of eight DAM methylation sites, GATC, was observed in a stretch of 245 bp that included the *E.coli* origin of replication (14). Is this a statistical cluster? We apply the formula [5] for $r = 7$, i.e., $Pr\{m^{(7)} > x/n^{8/7}\} \approx \exp\{-x^7/7!\}$ where *n* is the number of GATC sites. In the 1.4×10^6 bp of available *E.coli* sequences, the GATC frequency is .0044. Thus, we predict about $n = .0044\times4.7\times10^6 \approx 20680$ GATC sites over the whole genome. Now $\exp\{-x^7/7!\} = .99$ when $x = 1.75$. Hence, the threshold value for $m^{(7)}$ is $4.7\times10^6\,(1.75/n^{8/7}) \approx 96$ bp. Thus, for a random sequence of the composition and length of *E.coli*, a stretch *not* exceeding 96 bp that contains 8 occurrences of the GATC site is a statistically significant cluster at the 1% level.

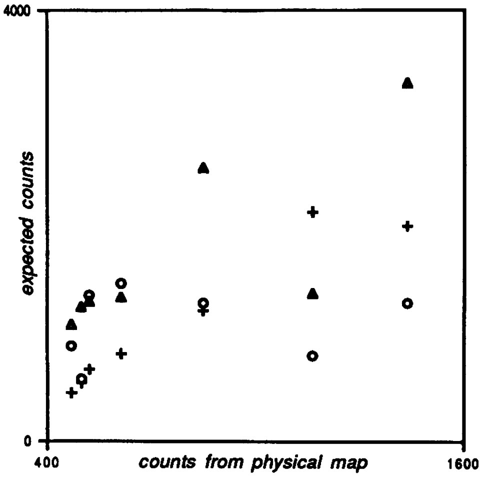

Figure 1: Comparison of observed and predicted site frequencies. The data of Table 2 are plotted showing the poor agreement between expected and observed counts of restriction sites. Enzymes can be identified from the physical map counts. Expected counts calculated from: (i) scaled sequence data, +; (ii) dinucleotide frequencies, ○; and (iii) trinucleotide frequencies, Δ.

Table 3. Summary of significant *r*-scans

		Minima			Maxima		
r	Enzyme	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1	EcoRV				64000	50400	41800
2	EcoRV				69400	67700	52500
10	PstI	9400	9700	10200			

The enzymes shown had significantly large *k*-maxima or significantly small *k*-minima *r*-scans, indicating overdispersion or clumping, respectively.

Table 4. Threshold values for *r*-scan minima

Number of Sites	*r*	Cutoff values (bp) 1%	 5%
500	1	<1	<1
	3	466	803
	5	2827	3917
1000	1	<<1	<1
	3	185	318
	5	1230	1705

For a representative number of restriction sites (500 or 1000) in a genome of length 4.7196×10^6 bp, minima of *r*-scans that are less than the tabled value may be declared significant, thus providing evidence of clumping.

The same formula shows that the presence of 8 GATC sites in a stretch of 245 bp would occur with probability about .06. So, as it stands, the observed concentration of GATC sites at ori-C of *E.coli* is not quite statistically significant. However, these sites in the ori-C segment may have value beyond the statistical evaluation, such as promoting accurate DNA replication with the help of DAM methylation and mismatch repair actions.

Overdispersion of NotI sites in chromosome 21? In a complete NotI (GCGGCCGC) digestion of chromosome 21, which has an estimated length of 50 Mb, a noncentromeric fragment approximately 5.4 Mb long was detected (personal communication of Charles Canter and Cassandra Smith). Is this unusually long? 'Postulating a random nucleotide distribution' and relying on formula [4] for the maximal fragment length ($Pr\{M^{(1)} < (\ln n + x)/n\} \approx \exp(-e^{-x})$, we deduce $M^{(1)} \geq$ 737.3 kb would be statistically significant at the 1% level. Thus, the 5.4 Mb gap between two consecutive NotI sites is indeed unusually long.

As always, statistical significance or lack thereof, does not provide a definitive statement about scientific significance. Statistical tests only provide benchmarks.

DISCUSSION

This paper sets forth statistical methods and interpretations concerning heterogeneity in DNA and protein sequences, exemplified principally by the counts and spacings of the Kohara *E.coli* physical map data. The Kohara map was constructed using partial digestion with seven 6-cutters (Table 1) in conjunction with the BglI-dyad symmetry cleavage site. Observations on counts and spacings of the Kohara map restriction sites are enigmatic. Specifically, there is substantial variability in counts among the different enzyme sites (Table 1), whereas the spacings of each type of site appear homogeneous, consonant with a random uniform distribution.

Counts

In the Kohara map there are several sources of error in the data. Especially, measurement errors were induced by recording restriction sites to the nearest 100 bp (rounding off). Stretches (about 10%) of the map were missing EcoRV sites, due to difficulties in reading autoradiographs. Moreover, the physical map appears to contain substantially fewer restriction sites than occur in the genome. Indeed, after screening about 1.43×10^6 bp of available cleaned *E.coli* sequences for restriction sites (13), and scaling appropriately, we estimate, with the exception of BamHI, that the *E.coli* genome contains between 12.7% (KpnI) and 86.4% (EcoRV) more sites than were mapped (Table 2). The sequence data revealed greater disparity for frequent cutters than for sparse cutters, strongly suggesting that the differences are largely due to undetected small intersite distances in construction of the physical map.

Predicted counts of sites based on binomial or Markov models using di- and trinucleotide frequencies (Table 2, Figure 1) corresponded poorly with observed counts and counts from extrapolated sequence data, indicating that restriction site distributions cannot be explained simply by nearest neighbor nucleotide interactions.

In light of these puzzling observations, it is natural to ask about the distribution of all 64 6 bp palindromes, 53 of which correspond to established restriction enzyme sites (15), in *E.coli* and in other organisms presumed to have co-evolved with *E.coli*, such as λ phage. In fact, about 60% of 6 bp palindromes in *E.coli* have substantially low counts (data not shown). The rare and frequent 6 bp palindromes largely coincide in *E.coli* and λ phage. It is interesting that the selection of enzymes used in (1) consists of the versatile mix of 3 dense, 1 average, and 3 rather sparsely distributed restriction sites.

Spacings

Extreme values of *r*-fragment lengths were used to detect significant clumping, overdispersion, or excessive regularity in the marker distribution. By varying *r*, organization on different scales can be detected, e.g., $r = 3$ can aptly detect near neighbor interactions while $r = 10$ can discern concentrations over a greater range. The *r*-fragment process is a moving sum process derived from the original first order process and so tends to smooth noisy fluctuations. Sums of *r* contiguous distances (*r*-fragment lengths) have a coefficient of variation (sample standard deviation divided by mean) inversely proportional to $\sqrt{r}$, rendering *r*-fragment lengths quite sensitive statistics in discerning clustering. The method of *r*-fragments was particularly useful for analyzing the *E.coli* data because of the digitation of restriction sites in units of 100 bp, precluding detection of clustering for $r = 1$ or 2, since for these models a minimum site separation of 0 bp is not unlikely.

At the level of restriction site spacings, the *E.coli* genome appears to be homogeneous (see Results) with two exceptions. The genome contains a striking cluster of PstI sites (based on 10-fragment lengths) and the least-frequent cutter, BamHI, tests to be overly evenly distributed. Since BamHI contains the DAM methylation site, GATC, there may be an adaptive purpose for such even spacings. Overdispersion of EcoRV sites occurred in the Kohara map, but this was an artifact of poor autoradiograph readings; the 1.43×10^6 bp *E.coli* sequence data did not reveal overdispersion for EcoRV locations (see Results and also Churchill *et al.* (3)).

CONCLUSION

Limitations of physical map data for characterizing genome organization are apparent. The effects of coarseness of the digitized data have been mentioned already. It is more difficult to isolate errors such as reversal of the order of closely-spaced sites or missing sites. Even without such problems, it would be erroneous to conclude, on the basis of an examination of the physical map alone, that the *E.coli* genome is 'homogeneous' at all levels of detail. In further examinations of the sequence data, a plethora of long direct and inverted repeats are found predominantly in intergenic regions spread approximately uniformly around the genome (16, 17). It is commonly known that the genomes of other organisms do not convey the degree of homogeneity apparent in the *E.coli* genome. For example, the spacings of λ-phage reveal two half-genomic sections, one G+C rich, the other A+T rich (established first from denaturization experiments (18)) with restriction enzyme sites roughly uniformly located within each region, but at unequal densities. Mammalian genomes exhibit analogous compartments (isochores), alternating approximately 100−200 kb of G+C and A+T rich regions (19). Other forms of inhomogeneity relate to CpG suppression, HTFII islands, polymorphic $(GT)_n$ repeats, *alu* dispersed elements, etc.

(20, 21). As more physical maps and sequence data come to hand, analytical methods such as those presented here can help in assessments of genome heterogeneity and organization.

ACKNOWLEDGEMENTS

We are grateful to Kenn Rudd for sharing data prior to publication. We thank C.Burge, C.Canter, and A.Campbell for valuable discussions on the manuscript. We also appreciate the helpful comments from two referees. This research was supported in part by NIH Grants HG00335-03 and GM10452-28 and NSF Grant DMS86-06244.

REFERENCES

1. Kohara,Y., Akiyama,K. and Isoro,K. (1987) *Cell* **50**, 495–508.
2. Rudd,K.E., Miller,W., Ostell,J. and Benson,D.A. (1990) *Nucl. Acids Res.* **18**, 313–321.
3. Churchill,G.A., Daniels,D.L. and Waterman,M.S. (1990) *Nucl. Acids Res.* **18**, 589–597.
4. Nussinov,R. (1984) *J. Mol. Evol.* **20**, 111–119.
5. Brendel,V., Beckmann,J.S. and Trifonov,E.N. (1986) *J. Biomol. Struct. Dynam.* **4**, 11–21.
6. McClelland,M., Jones,R., Patel,Y. and Nelson,M. (1987) *Nucl. Acids Res.* **15**, 5985–6005.
7. Stückle,E.E., Emmrich,C., Grob,U. and Nielsen,P.J. (1990) *Nucl. Acids Res.* **18**, 6641–6647.
8. Karlin,S. and Taylor,M.H. (1975) *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.
9. Cressie,N. (1977) *Austral. J. Statist.* **19**, 132–143.
10. Holst,L. (1980) *J. Appl. Probab.* **17**, 284–290.
11. Dembo,A. and Karlin,S. (1991) *Ann. of Appl. Prob.* in press.
12. Karlin,S. and Macken,C. (1991) *J. Amer. Statist. Assoc.* **86**, 26–33.
13. Rudd,K.E., Miller,W., Werner,C., Ostell,J., Tolstoshev,C. and Satterfield,S.G. (1991) *Nucl. Acids Res.* **19**, 637–647.
14. Krawiec,S. and Riley,M. (1990) *Microbiological Reviews* 502–539.
15. Roberts,R. (1990) Restriction Enzyme Database 9009.
16. Yang,Y. and Ames,C.F. (1990) In Drlica,K. and Riley,M. (eds), *The Bacterial Chromosome*. Amer. Soc. of Microbiology, Washington, DC, pp. 211–225.
17. Gilson,E., Saurin,W., Perrin,D., Bachellier,S. and Hofnung,M. (1991) *Nucl. Acids Res.* **19**, 1375–1383.
18. Inman,R.B. (1966) *J. Mol. Biol.* **18**, 464–472.
19. Bernardi,C., Mouchirond,D., Gautier,C. and Bernardi,G. (1988) *J. Mol. Evol.* **28**, 7–18.
20. Bird,A.P. (1986) *Nature*, **321**, 209–213.
21. Jukes,T.H. and Bhushan,V. (1986) *J. Mol. Evol.* **24**, 39–44.