STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:2409–2420 (DOI: 10.1002/sim.910)

Optimal estimation of transposition rates of insertion sequences for molecular epidemiology

Mark M. Tanaka[1,2,*,†] and Noah A. Rosenberg[1]

[1] *Department of Biological Sciences, Stanford University, CA 94305, U.S.A*
[2] *Department of Biology, Emory University, 1510 Clifton Road, Atlanta, GA 30322, U.S.A*

SUMMARY

Outbreaks of infectious disease can be confirmed by identifying clusters of DNA fingerprints among bacterial isolates from infected individuals. This procedure makes assumptions about the underlying properties of the genetic marker used for fingerprinting. In particular, it requires that each fingerprint changes sufficiently slowly within an individual that isolates from separate individuals infected by the same strain will exhibit similar or identical fingerprints. We propose a model for the probability that an individual's fingerprint will change over a given period of time. We use this model together with published data in order to estimate the fingerprint change rate for IS*6110* in human tuberculosis, obtaining a value of 0.0139 changes per copy per year. Although we focus on insertion sequences (IS), our method applies to other fingerprinting techniques such as pulsed-field gel electrophoresis (PFGE). We suggest sampling intervals that produce the least error in estimates of the fingerprint change rate, as well as sample sizes that achieve specified levels of error in the estimate. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Molecular technology for the identification of pathogenic strains has recently revealed important patterns of strain variation in infectious diseases such as tuberculosis [1, 2], salmonella [3, 4] and rice blight [5, 6]. Known as DNA fingerprints, molecular signatures of pathogenic strains are usually based on restriction fragment length polymorphisms. For instance, the technique of pulsed-field gel electrophoresis (PFGE) involves digesting a genome with a rare-cutting restriction enzyme and separating the fragments using appropriate conditions [7–9]. Another commonly used method for generating fingerprints utilizes the variation produced by insertion sequence (IS) elements in the genome of the pathogenic agents. With time, IS elements can shift locations, replicate into other genomic positions, and can be excised from the genome. Collectively, we refer to these changes in the number and positions of IS elements

[*] Correspondence to: Mark M. Tanaka, Department of Biology, Emory University, 1510 Clifton Road, Atlanta, Georgia 30322, U.S.A.
[†] E-mail: mmtanak@emory.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: GM28016, GM28428

Received August 2000
Accepted November 2000

Table I. A representative list of IS elements used for fingerprinting in molecular epidemiology.

Organism	Insertion sequence(s)	Reference
Bordetella spp.	IS*1001*	[27]
Lactobacillus helveticus	ISL2	[28]
Mycobacterium tuberculosis	IS*6110*	[1, 2, 10]
Mycoplasma incognitus	IS-like elements	[29]
Salmonella spp.	IS*200*	[30, 3, 4]
Shigella spp.	IS*1* and IS*2*	[31]
Streptomyces spp.	IS*1629*	[32]
Xanthomonas oryzae	IS*1112*	[5]
Yersinia spp.	IS*200*-like element	[33]

as transposition events. Polymorphism among strains results from variation in the number of elements, or the *copy number*, and their positions in the genome of the pathogen. Since they facilitate a relatively rapid and straightforward assessment of genotypes among strains of an infectious agent, IS elements are convenient genetic markers. Indeed, much recent research effort has been devoted to the identification and characterization of IS elements for fingerprinting a variety of bacterial pathogens (Table I).

Once IS elements have been identified for a pathogen of interest, studies can utilize IS-based polymorphisms to investigate potential epidemics [10]. The relatedness of fingerprint genotypes can be used as a supplement to conventional epidemiological techniques, such as contact tracing based on socio-demographic data, as a method of clustering infected individuals [1, 2]. At the population level, if a set of identified strains share the same or similar fingerprints, it is inferred that they derive from the same source, and that an outbreak of recent origin may be responsible. Clusters of related fingerprints are used in tuberculosis research, for example, since it is important to know the extent to which newly identified cases derive from reactivation of latent infections, compared to the amount due to recent disease transmission originating at a common source [2, 11–13]. This information can then be used to direct programmes for the control of the disease: should control strategies attempt to target modes of disease transmission or should they limit the reactivation of latent infections?

The use of molecular data to make assessments about whether a sample of isolates implies the existence of an epidemic requires fairly precise knowledge of the rate at which fingerprints change [14–16]. If the DNA marker changes very rapidly, isolates derived from the same source will be strongly differentiated, and without proper knowledge of the fingerprint change rate, the severity of outbreaks may be underestimated. If the marker changes very slowly, however, little variation in fingerprints will be observed, and knowledge of the change rate is necessary in order to avoid overestimation of the magnitude of an outbreak. With a slowly changing marker, if the observed polymorphism is great, the variation is likely to be old, and the disease is probably endemic at low frequency in the population. It is therefore crucial to verify that the fingerprint change rate of a marker is neither too high nor too low for use in molecular epidemiology.

Empirical estimates of transposition rates have been made for IS elements IS*1*, *2*, *3*, *4*, *5*, *30*, *150* and *186* in *Escherichia coli* [17]. Using a linear regression model and assuming that strains of all copy numbers undergo the same rate of change, Naas *et al.* [17] arrived at an estimate of 0.08 changes (gain or loss of bands) per year per culture, which, when divided by a typical copy number of 10, gives 0.008 changes per transposable element per culture per year. Although this method is effective, it can be improved by taking into account the copy number associated with each strain, since it is known that high-copy strains change more rapidly than low-copy strains (for example, see references [18, 16]).

Relatively fast changes in fingerprint patterns based on IS*6110* have been observed in *Mycobacterium tuberculosis* [19, 16], from which preliminary estimates of fingerprint change rates can be calculated. De Boer *et al.* [14] used survival analysis – modelling the survival of the identity of fingerprints over time as a process of exponential decay – to estimate the rate of IS-based fingerprint changes. Using IS*6110* data, they concluded that the half-life of an individual fingerprint was 3.2 years. The 95 per cent confidence interval for this figure corresponds to a rate of 0.0138–0.0330 changes per element per year if we assume that a typical copy number is 10. This method is useful, but it can be improved by taking copy number into account and by using sample intervals more efficiently. In their Figure 1, de Boer *et al.* [14] confirm that high copy strains tend to change faster than low copy strains, demonstrating that copy number could be incorporated in order to provide information about the fingerprint change rate. Additionally, actual sampling times could potentially be used in place of discretized values.

We propose here a systematic procedure for quantifying rates of fingerprint change, which incorporates the fact that strains of different copy number change at different rates. We distinguish *transposition rates* and *fingerprint change rates*. Transposition rates are usually reported in units of events per cell division in cultured bacteria [20]. We suggest a related but alternative parameter that is more directly applicable to epidemiological situations that involve within-host dynamics of pathogen populations. Fingerprint change rates describe the rate at which each band in a DNA fingerprint changes over time and they subsume all pathogen population processes within an individual [21, 22] that may affect the genetic identification of isolates. A fingerprint change within a host or culture results from a transposition event followed by replacement of a pre-existing strain by the transposed strain. We assume in our model that a fingerprint of an isolate indicates the presence of only a single strain within an individual; as fingerprinting techniques mature so that the relative frequencies of bacteria of multiple strains in an individual can be measured and understanding of within-host processes improves, our procedure should be revisited.

In this article, we propose an efficient method for estimating the rate of change of IS-based fingerprints. We also suggest how to optimize experimental design to extract maximal information about transposition rates from a full genetic data set.

2. OPTIMAL ESTIMATION OF CHANGE RATES

2.1. A model for changes in IS fingerprints

We model the number $N(t)$ of fingerprint changes in a given individual during time interval t as a Poisson process, with the rate of the process proportional to the individual's copy number k. This model assumes that copies mutate independently of each other and that differences among strains in the ability to proliferate and replace competing strains are not systematically related to copy number. We let θ be the rate of the process when the copy number is 1; that

Table II. The likelihood of fingerprint outcomes. The fingerprint for each typed individual behaves as a Bernoulli trial with parameter based on the copy number, the interval between fingerprints and the fingerprint change rate. First partial derivatives of $\ln f$, which are used in the derivation of optimal sampling intervals, are also given.

g	$f(g,k,t\|\theta)$	$\frac{\partial}{\partial\theta}\ln f(g,k,t\|\theta)$
0	$1-w(k,t,\theta)=\mathrm{e}^{-(k\theta)t}$	$-kt$
1	$w(k,t,\theta)=1-\mathrm{e}^{-(k\theta)t}$	$kt\mathrm{e}^{-k\theta t}/(1-\mathrm{e}^{-k\theta t})$

is, θ is the instantaneous fingerprint change rate measured in units of changes per copy per unit time (we use years for the time unit). Because transposition events are assumed to be independent of one another, the instantaneous rate of change of a k-copy strain is $k\theta$. Thus, for changes that do not affect copy number, the probability of j changes occurring in a k-copy strain over a time-span t would be

$$\Pr[N(t)=j]=\frac{(k\theta t)^j\mathrm{e}^{-k\theta t}}{j!} \tag{1}$$

However, experimental methods do not usually permit the exact number of transposition events that took place during the time period between two fingerprints. Therefore, we group all possible changes into a single class. This also circumvents the problem that $k\theta$ changes after the first change in copy number. We define $w(k,t,\theta)$ to be the probability of *at least one* change occurring in the fingerprint of a k-copy strain during time t:

$$w(k,t,\theta)=\Pr[N(t)\geqslant 1]=1-\mathrm{e}^{-k\theta t} \tag{2}$$

We let G be a Bernoulli trial which takes the value 1 with probability $\Pr[N(t)\geqslant 1]$. Table II indicates the likelihood function f of the measurements k and t for a single individual, together with the observation g of the random variable G, conditioned on the unknown parameter θ.

Although we use a specific model to estimate the rate of fingerprint change, the method can readily be modified to implement different models. For instance, we can include regulation of transposition [23]; one way to model this phenomenon is to set the transposition function $w(k,t,\theta)$ such that events become less likely as copy numbers increase beyond a certain value. Further parameters specifying the nature of additional features must then be jointly estimated.

2.2. Estimating the fingerprint change rate

Each isolate (individual), i, where $i=1,\dots,n$, and where n is the total number of isolates examined, is fingerprinted at two times that are separated by an intervening length of time t_i. Let k_i be the copy number of strain i in the initial fingerprint. By the time of the second fingerprinting, a change may have occurred in the fingerprint. This change represents a mutation in the IS elements of one pathogen followed by the replacement of the infectious strain by these mutant pathogens. We do not allow the possibility of multiple distinguishable fingerprints from a single isolate.

For each isolate, we measure three quantities: the observation or whether or not a change occurred (g); the copy number initially associated with the strain (k), and the time interval between the two samplings of the same isolate (t). The ith observation is associated with a vector $(g_i,\ k_i,\ t_i)$.

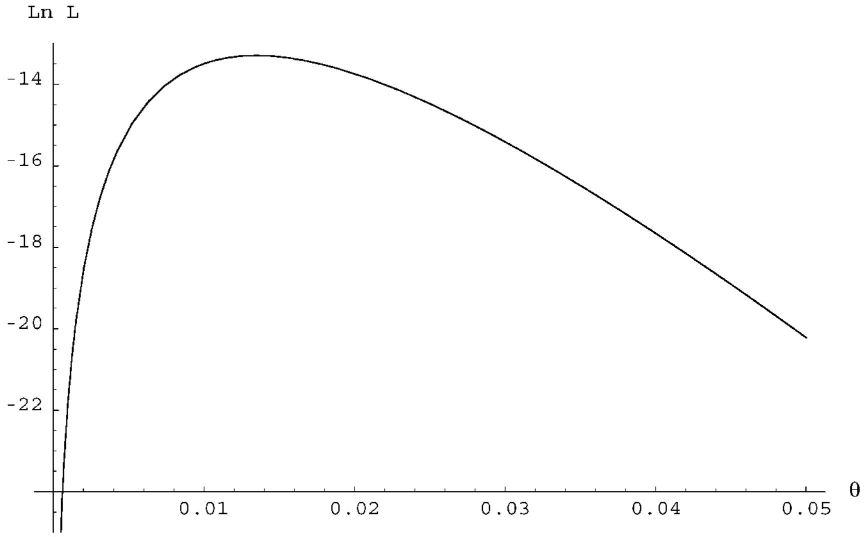

Figure 1. The log-likelihood curve (Ln L) for the parameter θ, using data from Niemann *et al.* [15].

The likelihood of the observed data set, in which the fingerprint change rate θ is unknown, is

$$\mathrm{Lik}(\theta)=\prod_{i=1}^{n} f(g_i,k_i,t_i|\theta) \tag{3}$$

Maximizing this likelihood with respect to θ gives rise to the maximum likelihood estimate (MLE), $\hat{\theta}$, of the fingerprint change rate.

Niemann *et al.* (reference [15], Table I) report a sample consisting of $n=56$ cases, using IS*6110* in *Mycobacterium tuberculosis*, of the appropriate nature for the application of our method. Numerical maximization of the likelihood of these data yields an estimate of $\hat{\theta}=0.0139$ per element per year for the fingerprint change rate of IS*6110*. We show the likelihood curve for these data in Figure 1. The curve is well-behaved, in that it has a single peak. An approximate standard deviation of the MLE, which derives from the asymptotic normality of maximum likelihood estimators, is

$$s_{\hat{\theta}}=\left(\sum_{i=1}^{n}\frac{(k_it_i)^2\mathrm{e}^{-k_i\hat{\theta}t_i}}{1-\mathrm{e}^{-k_i\hat{\theta}t_i}}\right)^{-\frac{1}{2}} \tag{4}$$

as discussed in the Appendix A. For the data set from Niemann *et al.* [15], a 95 per cent confidence interval for the MLE is (0.0028, 0.0251).

We note that the 95 per cent confidence interval of the fingerprint change rate derived from the estimates of de Boer *et al.* [14] is of comparable magnitude to ours. However, the greater efficiency of maximum likelihood methods together with our incorporation of copy number information into our mutation model allows us to achieve similar precision using a substantially smaller data set.

2.3. Optimal sampling

We suggest two ways in which the sampling protocol may be improved to address the precision of the maximum likelihood estimator. First, if the sampling time interval is chosen at the discretion of the experimenter, as may be the case for pathogens in laboratory or domesticated animals, we can find the optimal time yielding the least error in the estimate. Second, we find the sample size required to achieve a particular desired level of precision.

The variance of the maximum likelihood estimator, ($s^2_{\hat{\theta}}$), depends on copy numbers, sampling intervals, and the estimator of the fingerprint change rate, $\hat{\theta}$. By minimizing $s^2_{\hat{\theta}}$ with respect to the sampling interval t, we can find the *optimal sampling interval* for a strain with a given copy number. In Appendix A, we show that this optimal sampling interval is given approximately by

$$t_{\text{opt}}(k, \hat{\theta}) = \frac{\xi}{k\hat{\theta}} \tag{5}$$

where ξ is a constant whose value is approximately 1.5936. In this way, the optimal sampling time for each individual, associated with a strain of a given copy number, can be determined. When we later consider the sample sizes required for particular levels of precision, we must take into account the entire distribution of copy numbers rather than assuming a single copy number.

In order to check that the approximations involved in determining the optimal sampling interval are reasonable, we simulated samples of strains undergoing the stochastic transposition process described in Section 2.1. For each simulated sample, we used a given copy number (5, 10 or 15) for every strain, a particular sampling interval, and a sample size of 56, identical to that of Niemann *et al.* [15]. We used the value $\theta_0 = 0.0139$ for the fingerprint change rate. The MLE was calculated for each simulated sample, and 10 000 replicates were simulated for each combination of copy number and sampling interval. The standard deviation of the distribution of the MLE was computed over these replicates for each combination. Figure 2 shows the simulated standard deviations under different sampling intervals.

The optimal sampling intervals for copy numbers $k = 5$, 10 and 15 are, respectively, 22.93, 11.46 and 7.64 years. In separate simulations, we used these times to calculate the simulated optimal standard deviations, which can be compared to the theoretical value. Note that when using t_{opt}, the theoretical standard deviation of MLEs (4) is independent of copy number. Thus, with a given sample size, the minimum possible standard deviation should be equal for all copy numbers, with a given sample size. The simulated values are presented in Table III, and they lie close to the predicted values.

Figure 2 shows that the standard deviations are high when the sampling interval is much lower than the optimal sampling interval, and that they decrease as the sampling interval approaches t_{opt}. When the sampling interval is substantially higher than the optimal interval, occasionally the fingerprints of all individuals in the sample change during the time interval. If all the fingerprints change, the likelihood function is maximized when θ is the maximum of the allowed domain. We therefore do not report the irregular and high standard deviations for sampling intervals beyond the optimum. We also note that the longest time interval (10 years) used in Figure 2 is beyond limits of feasibility in empirical studies.

We now consider the sample size required to achieve a desired precision. The analysis of optimal sampling intervals above requires that the time of second sampling can be controlled.

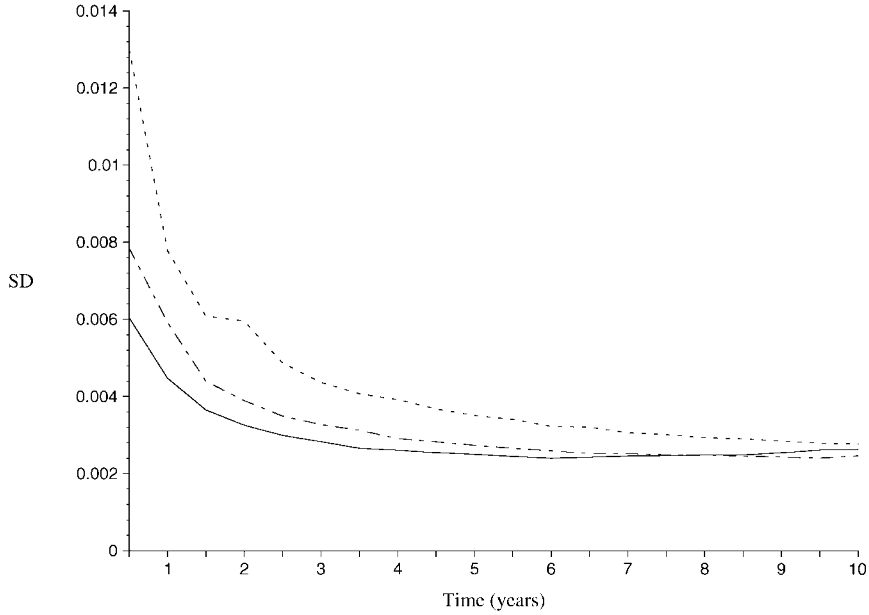

Figure 2. Standard deviations of simulated MLEs. Sample size $n=56$; $\theta=0.0139$; 10 000 replicates for each sampling time interval/copy-number. Solid line: $k=15$ copies; dot-dashed: $k=10$ copies; dashed: $k=5$ copies.

Table III. Simulated standard deviations, assuming $\theta=0.0139$, calculated from 10 000 replicates for each copy number. The theoretical standard deviation for $\theta=0.0139$, $n=56$ and any copy number is 0.002308.

Copies	t_{opt}	Simulated SD
5	22.93	0.00246
10	11.46	0.00245
15	7.64	0.00248

Because this will not always be the case, as with human diseases in which patients are sampled during hospital visits, we also examine the situation in which distributions of sampling intervals and copy numbers are given, and we ask what sample size is required in order to ascertain the change rate to a specified level of confidence. That is, we find the relationship between sample size and standard deviation of MLEs when empirical copy numbers and sampling time intervals are used. We again take advantage of a simulation approach, similar to that used above. The method is as follows.

The size of each simulated sample is taken from the set $\{40, 50, 60, \ldots, 190\}$. Each isolate is associated with a copy number drawn from a prespecified distribution of copy numbers, and a sampling interval drawn independently from a distribution of times. The empirical copy number distribution used here is taken from the set of fingerprints presented in Tanaka

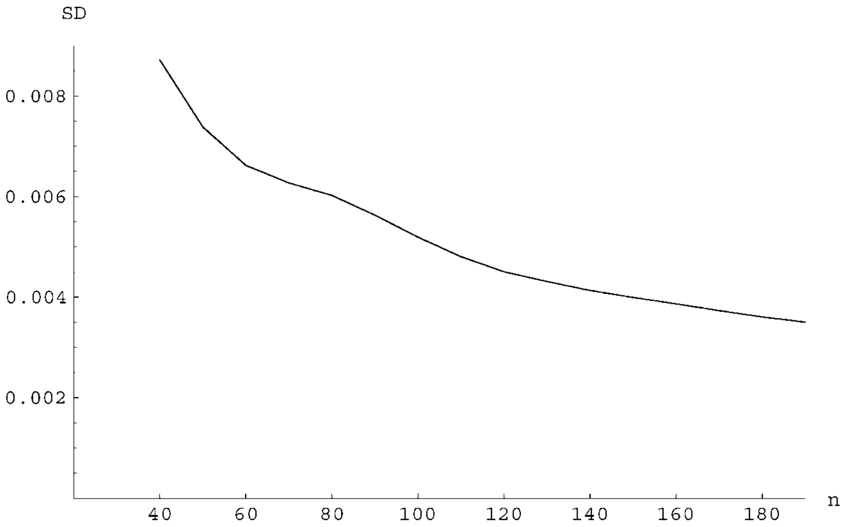

Figure 3. Standard deviation (SD) of MLE of fingerprint change rate as a function of sample size (n). For each sample size 100 000 replicates were simulated.

et al. [24]. The empirical distribution of sampling intervals is taken from Niemann *et al.* [15]. With $\theta = 0.0139$, we simulate the stochastic process described above, for each isolate. The MLE of the fingerprint change rate is calculated from the sample. For each sample size, 100 000 replicates were simulated, and the standard deviation of the MLEs calculated over these replicates.

Figure 3 plots the resulting standard deviations as a function of sample size. Thus, for example, for the standard deviation to be as low as 0.004, which is 29 per cent of the underlying value of θ, a sample size of $n = 150$ is required. We note that a more detailed consideration of this issue as a cost-benefit problem can yield a preferred sample size.

3. DISCUSSION

This article proposes an efficient method for estimating the rate of DNA fingerprint changes; the efficiency of the method stems from the fact that it makes use of the information contained in copy numbers and sampling times. In addition, we suggest ways of improving the experimental design by optimizing the sampling interval or increasing the sample size appropriately.

We note that the optimal sampling interval and $s_{\hat{\theta}}$ depend in interesting ways on the value of the fingerprint change rate and the copy numbers of the genotyped strains. For any copy number, at the optimal sampling interval, the fraction of fingerprints which have changed is approximately 0.7968, as given by the equation $w(k, t, \theta) = 1 - \mathrm{e}^{-k\theta t_{\mathrm{opt}}} = 1 - \mathrm{e}^{-k\theta(\xi/k\hat{\theta})} \simeq 1 - \mathrm{e}^{-\xi}$. The optimal sampling interval t_{opt} decreases as the copy number (k) increases. Since high-copy strains change faster than low-copy strains, it takes less time before this optimal fraction of fingerprints with high copy number have changed. A sampling protocol for estimating fingerprint change rates might therefore focus on high-copy individuals, since it takes less time for them to produce more precise rate estimates.

The optimal sampling interval also decreases as $\hat{\theta}$ increases. This relationship derives from the fact that higher transposition rates lead to faster changes and it takes less time for the optimal fraction of fingerprints to have changed.

The optimal sampling interval for IS elements will generally be very high, even when the fingerprint change rate is larger than that used in this study. For example, even with extreme values of $k = 30$ and $\theta = 0.05$, t_{opt} is about one year. Our study of the data of Niemann *et al.* [15] leads us to suggest that if different families of IS elements have relatively similar rates of change, long sampling intervals should also be used to measure fingerprint change rates for other IS elements besides IS*6110*.

For time-pressured epidemiological applications, we have identified a trade-off between obtaining fingerprint change rates quickly and obtaining them with precision. Our analysis suggests that if the goal of an experiment is to estimate the fingerprint change rate precisely, one should wait as long as practical considerations allow before taking the second sample. In practice, the maximal allowable time generally will not exceed the optimal sampling interval. However, a less precise estimate can be produced in a short amount of time, though we recommend intervals of at least one year when possible.

This approach is reasonable, as indicated by the rapid decay in the standard deviation in $\hat{\theta}$ with respect to t_{opt}, as the optimal value is approached, especially when copy numbers are high (Figure 2). When long time intervals are available to researchers, the sampling interval can be decided on the basis of individual fingerprint scores, so that the second fingerprints of different individuals are taken at different times.

The modelling framework we have used assumes a fairly simple transposition process. Extensions could incorporate, for example, regulation of transposition (or copy number control), temporally heterogeneous transposition rates, within-host selective sweeps dependent on copy number, or multiple transpositions with different transposition rates. The specific nature of the transposition process provides an equation for the probability, w, of a change during the sampling interval. With more complicated expressions for w, the rest of our method can proceed in the same way, by maximizing the likelihood of observations and then minimizing the sampling variance with respect to time.

If multiple parameters are jointly estimated, as would occur if for instance regulation were incorporated, the analysis would utilize analogous multidimensional results from maximum likelihood theory (see reference [25], for example). An optimality condition on the sampling interval can be obtained by minimizing an appropriate function of the variances of the parameters [26].

Extensions of the analysis are also possible with respect to the sampling regime; confidence intervals on the rate estimate can be decreased if individuals are fingerprinted more than twice. Thus, an optimal sampling scheme can be devised to identify multiple times at which to screen individuals.

Our methods can be generalized to molecular markers other than insertion sequences, for which different expressions for w can be proposed. Using a model for how changes take place over time, parameters representing the rate of change can be optimally inferred. With pulsed-field gel electrophoresis (PFGE), which is commonly used in molecular epidemiology for the purpose of genetic fingerprinting, our method can be applied directly without modification.

Assuming that new restriction sites do not appear, the estimated parameter will be the rate of change of each existing restriction site (in units of changes per site per year).

A useful consequence of the precise estimation of fingerprint change rates is that if more than a single family of IS elements have been characterized for a given bacterial pathogen, the most informative element for epidemiology can be selected for further use as a marker. For a marker to be useful in epidemiology, its rate of change must be commensurate with the rate at which the disease spreads. The exact criteria for commensurability remain to be determined.

4. CONCLUSIONS

We propose the following paradigm for the estimation of fingerprint change rates. Suitable individuals should be identified and fingerprinted at a first time ('time zero'). Based on the most recent estimate of the fingerprint change rate (which we provide here for *M. tuberculosis*/IS*6110*) and the copy numbers of the individuals, we recommend times at which those individuals should be fingerprinted a second time in order to minimize the variance of the re-estimated fingerprint change rate. If there is a practical upper limit on the time at which the second fingerprinting can occur, we recommend sample sizes that give a prespecified level of confidence about the rate. With the complete data set the likelihood method outlined in this paper may be used to compute the rate of change (maximizing equation (3) with respect to θ), with a confidence interval calculated from asymptotic results (using (4) as the standard deviation of the MLE).

It is important to note that a precise estimate of the fingerprint change rate in itself, while useful for understanding properties of the transpositions, is of primary interest for detection and control of epidemics. This change rate can also be incorporated in models treating epidemiology and genetics of the pathogen simultaneously, such as that of Tanaka *et al.* [24], in which a modestly precise estimate is adequate. Future studies of fingerprint change rates should also take into account the precision required for cost-efficient genotyping.

APPENDIX A: DERIVATION OF OPTIMAL SAMPLING INTERVALS

The large sample distribution of the maximum likelihood estimator $\hat{\theta}$ has an approximate variance of

$$s_{\hat{\theta}}^2 = \frac{1}{\sum I(k_i, t_i, \hat{\theta})} \tag{A1}$$

where

$$I(k_i, t_i, \theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln(f(g_i, k_i, t_i, \theta))\right)^2\right] \tag{A2}$$

(see, for example, reference [34]). Since maximum likelihood estimators are asymptotically normally distributed, an approximate 95 per cent confidence interval for $\hat{\theta}$ is $(\hat{\theta} - 1.96s_{\hat{\theta}}, \hat{\theta} +$

$1.96 s_{\hat{\theta}}$). In order to derive the optimal sampling interval for a given copy number k (let $k_i = k$ for all i) assume that $t_i = t$ for all i. Then for a sample of size n

$$s_{\hat{\theta}}^2 = \frac{1}{nI(k,t,\hat{\theta})}. \tag{A3}$$

By maximizing I with respect to t, we can find the time interval for a given copy number that minimizes the variance of the MLE. Using Table II and after some algebra, we obtain

$$I(k,t,\theta) = \frac{(kt)^2 \mathrm{e}^{-k\theta t}}{1 - \mathrm{e}^{-k\theta t}}. \tag{A4}$$

Now consider t such that $\frac{\partial I}{\partial t} = 0$. Restricting the domain of parameters to $\theta > 0$, $k > 0$ and $t > 0$, it can be shown that the optimal sampling interval, t_{opt}, is the non-zero solution of

$$1 - \mathrm{e}^{-k\hat{\theta}t} - \frac{k\hat{\theta}t}{2} = 0. \tag{A5}$$

The form of this transcendental expression suggests the solution $t_{\mathrm{opt}} = \xi/(k\hat{\theta})$, where ξ is the non-zero constant which satisfies $1 - \mathrm{e}^{-\xi} - \xi/2 = 0$. Numerically solving this equation, $\xi \simeq 1.593624$. It can easily be shown that $\xi/(k\hat{\theta})$ is the only non-zero solution and that it maximizes rather than minimizes $I(k,t,\hat{\theta})$.

ACKNOWLEDGEMENTS

We are grateful to A. Hirsh, M. Feldman, B. Levin, M. Lipsitch, S. Ptak, P. Small, J. Storey and the reviewers for their comments on this work. N. A. R. is supported by a National Defense Science and Engineering Graduate Fellowship. This research was funded by National Institutes of Health grants GM28016 and GM28428 to Marcus Feldman.

REFERENCES

1. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK. The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. *New England Journal of Medicine* 1994; **330**:1703–1709.
2. van Soolingen D, Hermans PWM. Epidemiology of tuberculosis by DNA-fingerprinting. *European Respiratory Journal* 1995; **8**:S649–S656.
3. Pelkonen S, Romppanen EL, Siitonen A, Pelkonen J. Differentiation of Salmonella serovar infantis isolates from human and animal sources by fingerprinting IS*200* and 16S *rrn* loci. *Journal of Clinical Microbiology* 1994; **32**:2128–2133.
4. Stanley J, Baquar N, Threlfall EJ. Genotypes and phylogenetic-relationships of *Salmonella typhimurium* are defined by molecular fingerprinting of IS*200* and 16S *rrn* loci. *Journal of General Microbiology* 1993; **139**: 1133–1140.
5. Adhikari TB, Vera-Cruz CM, Zhang Q, Nelson RJ, Skinner DZ, Mew TW, Leach JE. Genetic diversity of *Xanthomonas oryzae* pv. oryzae in Asia. *Applied and Environmental Microbiology* 1995; **61**:966–971.
6. Vera-Cruz CM, Ardales EY, Skinner DZ, Talag J, Nelson RJ, Louws FJ, Leung H, Mew TW, Leach JE. Measurement of haplotypic variation in *Xanthomonas oryzae* within a single field by rep-PCR and RFLP analyses. *Phytopathology* 1996; **86**:1352–1359.
7. Johnson JM, Weagant SD, Jinneman KC, Bryant JL. Use of pulsed-field gel electrophoresis for epidemiologic study of *Escherichia coli* O157:H7 during a food-borne outbreak. *Applied and Environmental Microbiology* 1995; **61**:2806–2808.
8. Krause U, Thomson-Carter FM, Pennington TH. Molecular epidemiology of *Escherichia coli* O157:H7 by pulsed-field gel electrophoresis and comparison with that by bacteriophage typing. *Journal of Clinical Microbiology* 1996; **34**:959–961.

9. Steffens L, Franke S, Nickel S, Schwarzkopf A, Flemmig TF, Karch H. DNA-fingerprinting of *Eikenella corrodens* by pulsed-field gel electrophoresis. *Oral Microbiology and Immunology* 1994; **9**:95–98.
10. van Soolingen D. Utility of molecular epidemiology of tuberculosis. *European Respiratory Journal* 1998; **11**:795–797.
11. Behr MA, Hopewell PC, Paz EA, Kawamura LM, Schecter GF, Small PM. Predictive value of contact investigation for identifying recent transmission of *Mycobacterium tuberculosis. American Journal of Respiratory and Critical Care Medicine* 1998; **158**:465–469.
12. Burman WJ, Reves RR, Hawkes AP, Rietmeijer CA, Yang ZH, El-Hajj H, Bates JH, Cave MD. DNA fingerprinting with two probes decreases clustering of *Mycobacterium tuberculosis. American Journal of Respiratory and Critical Care Medicine* 1997; **155**:1140–1146.
13. Rhee JT, Tanaka MM, Behr MA, Agasino CB, Paz EA, Hopewell PC, Small PM. Use of multiple markers in population-based molecular epidemiologic studies of tuberculosis. *International Journal of Tuberculosis and Lung Disease* 2000; **4**:1111–1119.
14. de Boer AS, Borgdorff MW, de Haas PEW, Nagelkerke NJD, van Embden JDA, van Soolingen D. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. *Journal of Infectious Diseases* 1999; **180**:1238–1244.
15. Niemann S, Richter E, Rusch-Gerdes S. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. *Journal of Clinical Microbiology* 1999; **37**:409–412.
16. Yeh RW, de Leon AP, Agasino CB, Hahn JA, Daley CL, Hopewell PC, Small PM. Stability of *Mycobacterium tuberculosis* DNA genotypes. *Journal of Infectious Diseases* 1998; **177**:1107–1111.
17. Naas T, Blot M, Fitch WM, Arber W. Dynamics of IS-related genetic rearrangements in resting *Escherichia-coli* K-12. *Molecular Biology and Evolution* 1995; **12**:198–207.
18. van Soolingen D, de Haas PEW, Hermans PWM, Groenen PMA, van Embden JDA. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis. Journal of Clinical Microbiology* 1993; **31**:1987–1995.
19. Alito A, Morcillo N, Scipioni S, Dolmann A, Romano MI, Cataldi A, van Soolingen D. The IS*6110* restriction fragment length polymorphism in particular multidrug-resistant *Mycobacterium tuberculosis* strains may evolve too fast for reliable use in outbreak investigation. *Journal of Clinical Microbiology* 1999; **37**:788–791.
20. Kleckner N. Transposable elements in prokaryotes. *Annual Review of Genetics* 1981; **15**:341–404.
21. Agur Z, Abiri D, van der Ploeg LHT. Ordered appearance of antigenic variants of African trypanosomes explained in a mathematical model based on a stochastic switch process and immune-selection against putative switch intermediates. *Proceedings of the National Academy of Sciences USA* 1989; **86**:9626–9630.
22. Antia R, Levin BR, May RM. Within-host population dynamics and the evolution and maintenance of microparasite virulence. *American Naturalist* 1994; **144**:457–472.
23. Kleckner N. Regulation of transposition in bacteria. *Annual Review of Cell Biology* 1990; **6**:297–327.
24. Tanaka MM, Small PM, Salamon H, Feldman MW. The dynamics of repeated elements: Applications to the epidemiology of tuberculosis. *Proceedings of the National Academy of Sciences USA* 2000; **97**:3532–3537.
25. Kendall M, Stewart A. *Advanced Theory of Statistics*, volume 2, 4th edn. MacMillan Publishing Co.: New York, 1979.
26. Atkinson AC, Donev AN. *Optimum Experimental Designs*. Oxford University Press: Oxford, 1992.
27. van der Zee A, Mooi F, van Embden J, Musser J. Molecular evolution and host adaptation of *Boredetella spp*: Phylogenetic analysis using multilocus enzyme electrophoresis and typing with 3 insertion sequences. *Journal of Bacteriology* 1997; **179**:6609–6617.
28. Zwahlen MC, Mollet B. ISL2, a new mobile genetic element in *Lactobacillus helveticus*. *Molecular and General Genetics* 1994; **245**:334–338.
29. Hu WS, Hayes MM, Wang RYH, Shih JWK, Lo SC. High-frequency DNA rearrangements in the chromosomes of clinically isolated *Mycoplasma fermantans*. *Current Microbiology* 1998; **37**:1–5.
30. Burr MD, Josephson KL, Pepper IL. An evaluation of DNA-based methodologies for subtyping Salmonella. *Critical Reviews in Environmental Science and Technology* 1998; **28**:283–323.
31. Soldati L, Piffaretti JC. Molecular typing of Shigella strains using pulsed field gel electrophoresis and genome hybridization with insertion sequences. *Research in Microbiology* 1991; **142**:489–498.
32. Healy FG, Bukhalid RA, Loria R. Characterization of an insertion sequence element associated with genetically diverse plant pathogenic *Streptomyces spp*. *Journal of Bacteriology* 1999; **181**:1562–1568.
33. Odaert M, Berche P, Simonet M. Molecular typing of *Yersinia pseudotuberculosis* by using an IS200-like element. *Journal of Clinical Microbiology* 1996; **34**:2231–2235.
34. Rice JA. *Mathematical Statistics and Data Analysis*. Wadsworth and Brooks/Cole: Belmont, California, 1988.