Statistical Science
2000, Vol. 15, No. 1, 27–45

Comparing DNA Fingerprints of Infectious Organisms

Mark R. Segal, Hugh Salamon and Peter M. Small

Abstract. Genotypes of infectious organisms are becoming the foundation for epidemiologic studies of infectious disease. Central to the use of such data is a means for comparing genotypes. We develop methods for this purpose in the context of DNA fingerprint genotyping of tuberculosis, but our approach is applicable to many fingerprint-based genotyping systems and/or organisms. Data available on replicate (laboratory) strains here reveal that (i) error in fingerprint band size is proportional to band size and (ii) errors are positively correlated within a fingerprint. Comparison (or matching) scores computed to account for this error structure need to be "standardized" in order to properly rank the comparisons. We demonstrate the utility of using extreme value distributions to effect such standardization. Several estimation issues for the extreme value parameters are discussed, including a lack of robustness of (approximate) maximum likelihood estimates. Interesting findings to emerge from examination of quantiles of standardized matching scores include (i) formal significance is not attainable when querying a database for a given fingerprint pattern and (ii) maximal matching probabilities are not necessarily monotonely decreasing with increasing numbers of fingerprint bands.

Key words and phrases: Extreme value distribution, genotyping, maximum likelihood estimation, moment estimation, tuberculosis.

1. INTRODUCTION

In recent years considerable attention has been given to new, emerging and reemerging infectious diseases (Institute of Medicine, 1992). The global epidemic of human immunodeficiency virus (HIV) demonstrates the potential impact of newly emerged pathogens. Simultaneously, old pathogens such as those which cause tuberculosis, cholera, plague, dengue and yellow fever are having a profound impact in various localities. In addition, mutation and selection are leading to drug-resistant strains of *Mycobacterium tuberculosis*, enterobacteria, malaria, pneumococci, gonococci and staphylococci.

Surveillance and applied research are two of the four critical goals identified by the Centers for Disease Control and Prevention (CDC) (1994) in their response to emerging pathogens. Surveillance emphasizes the detection, monitoring and investigation of infectious diseases. These activities include conventional epidemiologic practices of tracking trends in particular species, such as rates of gonorrhea in a population and rates of drug resistance among these cases. Applied research, which integrates the principles and practices of molecular biology and population genetics into these efforts, may provide even greater insights. For example, by using molecular epidemiologic approaches it is possible to determine if drug resistance is simultaneously emerging in numerous different strains (suggesting a need to modify antibiotic utilization) or from the clonal dissemination of a single strain (suggesting that efforts to interrupt disease

Mark R. Segal is Professor, Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-0560 (e-mail: mark@biostat.ucsf.edu). Hugh Salamon is Researcher, Berlex Laboratories, Inc., 15049 San Pablo Avenue, Richmond, California 94804-0099. Peter M. Small is Assistant Professor, Division of Infectious Diseases and Geographic Medicine, Stanford University Medical Center, Stanford, California 94305-5107.

transmission should be intensified). Hence, the CDC has specifically identified "the expanded use of molecular epidemiology in investigating emerging disease" as a central part of its response to emerging pathogens.

1.1 DNA Fingerprints of Infectious Organisms: *IS6110* Genotyping of Tuberculosis

The vast majority of the molecular data collected in molecular epidemiologic studies of infectious organisms take the visual form of DNA "fingerprints." Although these fingerprints can be generated through a variety of techniques, we will focus on one: *restriction fragment length polymorphism*. We will also concentrate on a single infectious organism, *M. tuberculosis*. Other organisms for which typing schemes have been devised, and for which databases of corresponding genotypes have been assembled, can potentially be analyzed analogously.

Recently, a technique has been developed that, for the first time, provides a practical method to identify and track specific *M. tuberculosis* strains. This genotyping technique is based upon the presence of a genetic sequence known as *IS6110*, a 1,355 base-pair (bp) repetitive element found in variable numbers and locations throughout the genome (Figure 1). This variability is exploited to generate strain-specific "DNA fingerprints," that is, to genotype the strain. In brief, DNA is extracted from the bacterial organism and cleaved into fragments using restriction enzymes (PvuII in Figure 1). These fragments are then electrophoretically separated according to size and transferred to nylon membranes, which are probed to detect those fragments containing *IS6110* (Figure 1). The resulting pattern of fragment or band sizes is called *a restriction fragment length polymorphism* (RFLP) pattern and is commonly known as a *DNA fingerprint*. Figure 2 displays a gel image of such patterns that features fingerprints (the columns or "lanes" as numbered at the top) corresponding to different specimens. Replicate H37Rv laboratory strains (see Section 2) appear on the leftmost and rightmost columns of this figure. A standard protocol for *IS6110* typing of *M. tuberculosis* is described in van Embden et al. (1993) and additional description is provided in Salamon, Segal and Small (1998).

The use of fingerprint data for molecular epidemiology purposes has, to date, largely consisted of treating sets of identical or "highly" similar fingerprints as indicating transmission of infectious disease from a common source. Thus, fingerprint comparisons are used to decide whether or not putative epidemiologic links exist between hosts of the sampled infectious organisms. These links

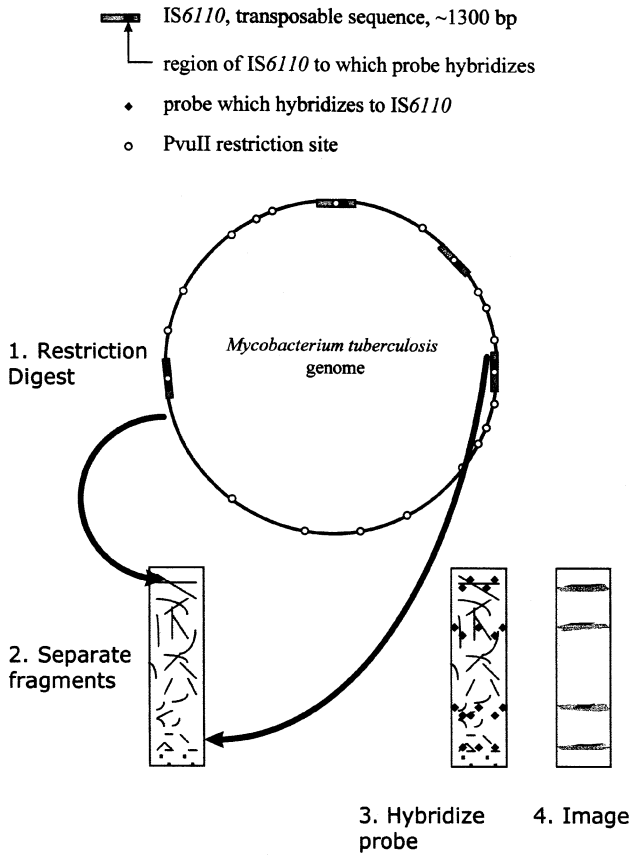

FIG. 1. *Restriction sites (circles) at which a restriction enzyme (PvuII) cuts ("digests") the genomic DNA are scattered throughout the genome: each IS6110 element (rectangles) contains such a restriction site, as well as a region to which a probe hybridizes or binds (dark squares). Thus, digesting the genomic DNA* (1) *results in DNA fragments of varying lengths, some of which contain the IS6110 element. These fragments are then separated electrophoretically* (2)*. Subsequently, a probe bound to a light-emitting enzyme (diamonds) is hybridized to those fragments* (3) *containing IS6110. The resultant light pattern corresponds to the lengths of fragments (band sizes) containing IS6110 and is recorded on photographic X-ray film* (4) *and later scanned into a computer image file.*

are further used to evaluate the extent of recent transmission and the emergence of drug-resistant or virulent strains. This program is only plausible given a background of diverse fingerprints.

To properly differentiate fingerprint comparisons we need to operationalize what constitutes "highly" similar fingerprints. Such assessments make recourse to band-matching algorithms that yield a matching score as detailed later (Section 2). Basically, such algorithms use error characteristics of the fingerprints to effect alignment of a fingerprint pair, with the subsequent count of corresponding bands (i.e., equally sized fragments) defining the matching score. However, because of the obvious dependence of such raw counts on the numbers

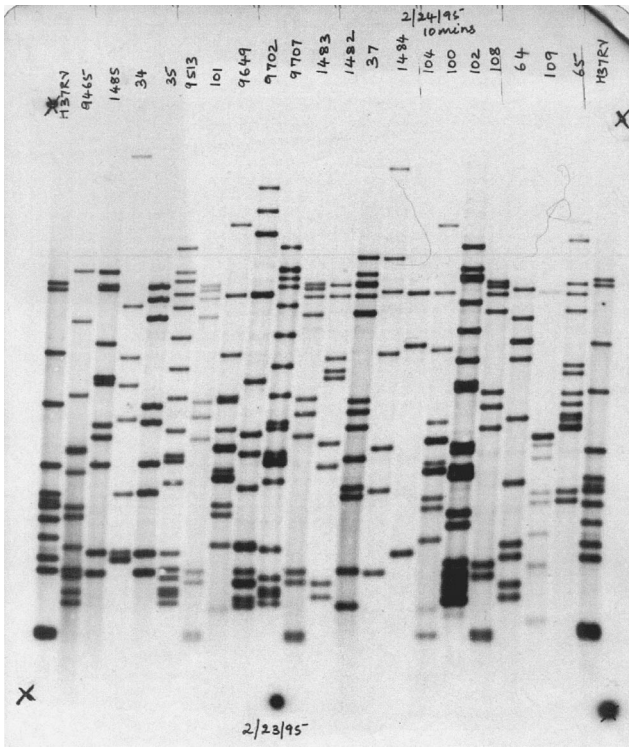

FIG. 2. *A gel image depicting IS6110 fingerprints of differing tuberculosis strains: the far left and right fingerprints correspond to replicates of the H37Rv laboratory strain.*

of bands in the fingerprints being compared, it is necessary to standardize these matching scores.

This need becomes apparent from consideration of identical or perfectly matched fingerprints. In essence, a fingerprint is a vector of band sizes. The length of the vector is the number of bands in the fingerprint. The components of the vector (i.e., band sizes) are estimated using reference markers of known size run on the same gel. A perfect match entails that the number and sizes of bands in both fingerprints coincide. With simple similarity measures, such as the Dice coefficient (which is just the proportion of matching bands), there is no differentiation between perfect agreement between two 5-, two 10-, or two 20-band fingerprints. This is contrary to the varying information content such perfect agreement represents.

Trying to effect standardization using theoretic models of the evolutionary processes underlying the fingerprints themselves is problematic because of the complexity of these processes. Beyond this, each individual fingerprinting technology would require its own specialized model development. Accordingly, we develop an empiric approach that provides a general prescription for standardizing matching scores.

Fingerprint data are being used extensively in part because of the advantages they have over other forms of genotyping. For example, relative to DNA sequence data, fingerprints are technically simple and inexpensive. Consequently, they can be used in epidemiologic studies with large sample sizes. Fingerprints can also simultaneously sample from throughout the organism's genome and so assess genetic variation at a variety of genomic locations. Finally, use of fingerprint data avoids concerns surrounding DNA sequence alignment.

However, each of these advantages comes at a cost. The technical simplicity results in several sources of experimental error which, in part, spawn a different set of alignment difficulties. The sensitivity to a multiplicity of genetic variations makes it impossible to identify specific genetic mechanisms underlying different fingerprint patterns. And, of foremost interest here, analytical approaches to this type of data are relatively primitive. This paper seeks to develop methodology for comparing fingerprints in the face of experimental error, as well as for standardizing the resultant matching scores. We next provide contrasts with the more developed area of DNA fingerprinting for forensic purposes, which demonstrates the need to devise new approaches for fingerprinting infectious organisms.

1.2 Contrast with Forensic Uses of DNA Fingerprints

As noted, to date there has been little investigation of statistical methods for infectious disease fingerprint data. This contrasts with the considerable attention directed toward the nominally related field of *human* DNA fingerprinting for forensic science and as evidence in criminal cases; see, for example, Roeder (1994).

However, the applicability of forensics (F) findings to molecular epidemiology (ME) of emerging pathogens is limited for the following reasons:

(i) The use of multiple single-locus probes (F) contrasts with the use of a single multiple-locus probe (ME). In a number of infectious disease fingerprinting schemes, including that used to type *M. tuberculosis* repetitive genetic elements found at multiple locations throughout the genome are exploited; see Figure 1.

(ii) A huge number of individual genotypes (essentially all are unique) (F) contrasts with a much smaller number of genotypes with many members (ME).

(iii) The genetics for a diploid organism (two sets of complete genomes per organism) with sex and

other recombination events decreasing linkage disequilibrium (i.e., weak association between genetic variants at different genomic locations) (F) contrasts with a haploid (single genome copy) system with little or no recombination resulting in a clonal population structure and hence a very high level of linkage disequilibrium (strong associations between genetic variants at different locations) (ME).

(iv) There exist differing ascertainment concerns.

(v) Interest solely in matching (F) raises issues distinct from those raised by interest in clusters and relatedness of patterns (ME).

Several of these items make analysis more complex in the molecular epidemiology setting. Still, a gel is a gel, and so commonalities with regard to experimental error exist as noted in Section 2.

1.3 San Francisco Database and Global Migration

During the 1990s a concerted effort has been made to study tuberculosis in San Francisco. Corresponding databases have been assembled. These data include extensive epidemiologic (e.g., age, gender, race, country of birth) and clinical (e.g., organs involved, therapy, x-ray, culture and AFB smear results, HIV status, antimicrobial susceptibility) data on virtually all 1,874 microbiologically confirmed cases of tuberculosis reported in San Francisco during the years 1991–1996. We have collected and performed RFLP analysis on 1,577 (84%) of these cases. Of these, we analyze here the 1,335 (85%) fingerprints dating from mid-1992, laboratory techniques being less stable prior to this date.

Our first comparisons of *IS6110* based RFLP patterns used Whole Band Analyzer (Genomic Solutions, Ann Arbor, Michigan), a commercially available UNIX-based system, to digitize and store these patterns. This system, which did not utilize information on band size errors, was used to identify similar patterns which must then be visually compared to determine identical matches (Woellfer, Bradford, Paz and Small, 1995). This is extremely laborious. Further, we have now exceeded the logistical limit of this approach and are unable to perform comparisons between large data sets, such as comparing all strains from San Francisco and Latin America.

Indeed, the focus of molecular epidemiological studies has expanded from studies of the transmission of *M. tuberculosis* in small, local outbreaks to evaluating its global migration. As outlined by Small (1995), this shift has resulted from the coupling of substantial geographic disparities in both overall tuberculosis case rates and multidrug-resistant tuberculosis with the dramatically increased mobility of the world's population. Public health prevention efforts require data on the nature and extent of disease spread. Consequently, the technical challenge has shifted from the comparison of a few DNA fingerprints to the analysis of thousands of patterns. Accordingly, a scoring algorithm has been developed (Section 2) to automate pattern comparisons. In Section 3 we describe approaches for standardizing these scores. Section 4 presents results of applying these methods to the San Francisco data and also makes external comparisons with fingerprints from Orizaba, Mexico. Section 5 offers concluding discussion.

2. AUTOMATED ALIGNMENT AND SCORING

We briefly describe some salient features of an automated algorithm for effecting fingerprint comparison; details are given in Salamon, Segal and Small (1998). The starting point for algorithm development was investigation of errors in fingerprint data. An empirical investigation of a laboratory tuberculosis strain (12-banded H37Rv), for which more than 100 replicate fingerprints were available, revealed the following:

1. Between-gel comparisons are subject to larger errors than within-gel comparisons (Figure 3), as would be anticipated.
2. For band sizes less than or equal to 5 kilobases (kb), error is proportional to band size (Figure 3). Such band sizes comprise 90% of the *M. tuberculosis* bands.

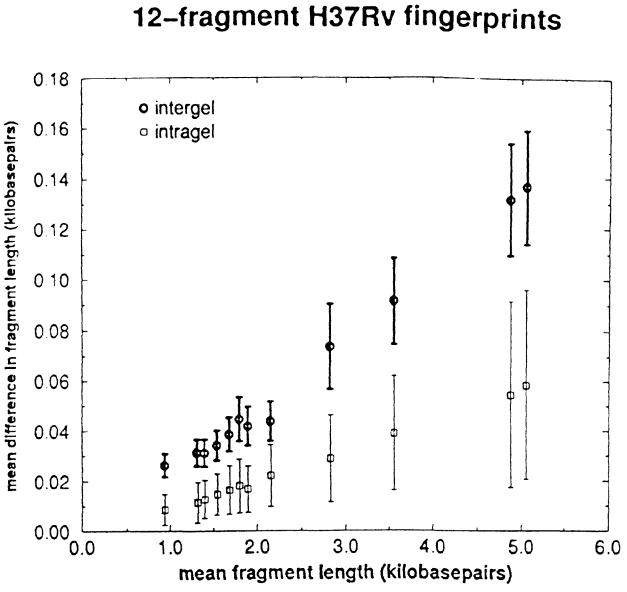

FIG. 3. *Error in band size is proportional to band size: the intervals around the means are* ± 2 *standard errors.*

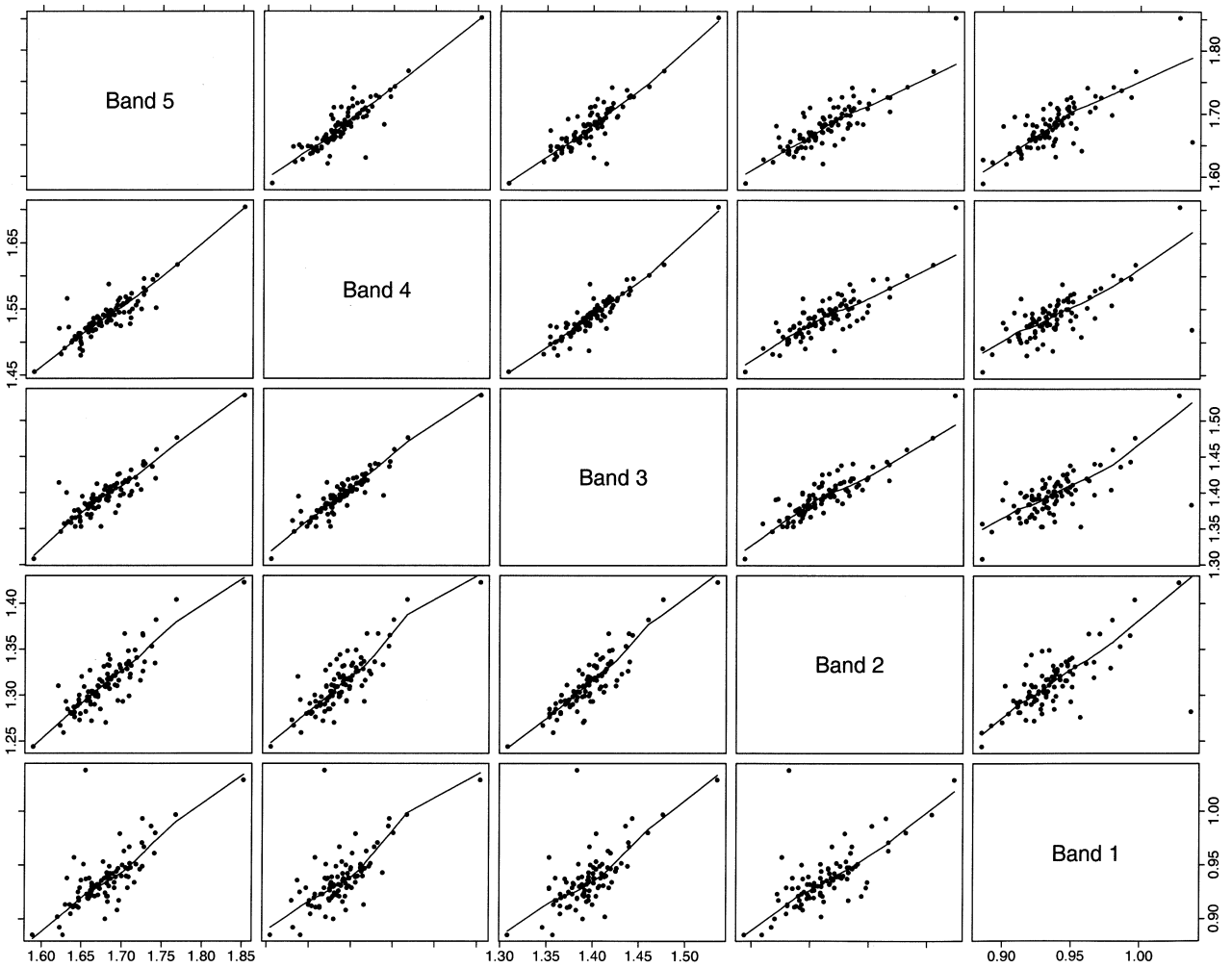

FIG. 4. *Smoothed scatterplot matrix showing strong positive dependence among the five smallest H37Rv bands.*

3. Error in one band is highly, and positively, correlated with error in measuring other bands in the same fingerprint; for example, if band 1 of a particular fingerprint is larger than average for H37Rv, then the other bands also tend to be larger than average. This positive correlation is displayed for the five smallest H37Rv bands in Figure 4.

These findings have also been documented in the forensic fingerprinting literature (Sudbury, Marinopoulos and Gunn, 1993; Eriksen and Svensmark, 1993).

That we observe proportional error together with correlated band size errors invites alignment of pairs of fingerprints by initially scaling the set of band sizes of one fingerprint to those of the other. The matching score is then obtained by counting the number of mutually closest bands within a prescribed (proportional) size difference. This approach requires two input parameters: (i) a range of scaling limit which restricts how much one fingerprint's bands can be inflated with respect to another fingerprint and (ii) a proportional error threshold for decreeing scaled bands to match or not. As briefly indicated below, the replicate patterns provide a basis for specifying these parameters. The applicability of the algorithm hinges on the above error characteristics. These require empiric assessment for a given combination of typing system and organism. They are not inherently determined by the biology of the organism.

For *IS6110* typing of *M. tuberculosis*, the replicates derive from running each gel with two or three H37Rv samples. Let the total number of replicates be r. All $N = r(r-1)/2$ pairwise comparisons of these replicates are performed, noting whether comparisons are between or within gels. Being replicates, we assume that each fingerprint has the same

number of bands, B. Let $m_{i,b}$ be the size (in kilobases) of band b (assigned from smallest to largest) of fingerprint i. Alignment of two replicate fingerprint patterns is effected using the following least squares criterion. We minimize

$$Q(\beta_{ij}) = \sum_{b=1}^{B} \left(\log(m_{i,b}) - \log(\beta_{ij} * m_{j,b})\right)^2$$

with respect to the scaling factor β_{ij} by which fingerprint j is aligned to fingerprint i. Log-transformed band sizes are used to reflect the fact that measurement error is proportional to band size. It is not critical whether scaling is applied before (as here) or after log transformation. The readily computable, closed-form solution for β_{ij} is

$$\hat{\beta}_{ij} = \exp\left(\frac{1}{B}\sum_{b=1}^{B} \log\left(\frac{m_{i,b}}{m_{j,b}}\right)\right).$$

Denote the scaled band sizes as m^*: if fingerprint j is scaled to fingerprint i then $m^*_{i,b} = m_{i,b}$ and $m^*_{j,b} = \beta_{ij} * m_{j,b}$. Comparing scaled fingerprints i and j we calculate band-specific absolute and proportional differences:

$$d_{i,j}(b) = |m^*_{i,b} - m^*_{j,b}|$$

and

$$r_{i,j}(b) = \frac{d_{i,j}(b)}{(m^*_{i,b} + m^*_{j,b})/2}.$$

The distribution of corresponding means (e.g., $\overline{r(b)} = (1/N)\sum_{i=1}^{r}\sum_{j>i}^{r} r_{i,j}(b)$) over all pairwise comparisons is used for setting error tolerances for declaring specific bands b to match as described in Salamon, Segal and Small (1998). Throughout, consideration is given to whether comparisons are between or within gels.

We now can discuss how to compare fingerprints possessing differing numbers of bands. The distribution of fitted scaling factors ($\hat{\beta}_{ij}$'s) (again over all pairwise comparisons) is used to establish the range for which alignment is attempted. Such alignment is effected by incrementally "sliding" fingerprints past one another and, in accord with the tolerances, counting the number of band matches. Setting the number of increments to 100 gives results closely agreeing with visual matching. The reported matching score is the maximum count over the 100 comparisons.

Numeric results and further details of the algorithm, including threshold settings that accommodate bands larger than 5 kb are given in Salamon, Segal and Small (1998). This "align-and-count" algorithm has proven invaluable for avoiding labor-intensive visual inspection and allows in excess of 50 pairwise comparisons of *IS6110* fingerprints per second on a SparcStation 20 or on low-end Pentium-based Unix platforms. The resulting comparisons are consistent with the labor-intensive matching in the San Francisco data. Furthermore, the align-and-count method has allowed comparison of San Francisco fingerprints with fingerprints from other localities, as illustrated in Section 4.

3. STANDARDIZING FINGERPRINT MATCHES

Application of the align-and-count algorithm yields a score: in comparing a fingerprint with m bands to another with n bands we observe $S = S_{m,n}$ matches. How we interpret and contrast these scores is important because, for example, when appropriately standardized they provide similarity measures which, in turn, serve as inputs to clustering algorithms used in tracing disease transmission. Questions of the following flavor arise:

- What is the probability, due to chance, that we would observe 8 matches when comparing two 8-banded patterns?
- Is it more surprising to observe 11 matches from a comparison of 13- and 15-band fingerprints than 7 matches from comparing 8- and 9-band fingerprints?

Addressing these questions requires distributional results for the matching scores. This will be the case regardless of the algorithm used to calculate such scores or, indeed, the nature of the genotyping system employed and the organism being studied. Thus, the approach to follow has much wider applicability than solely align-and-count scoring for *IS6110* typing of tuberculosis.

Directly obtaining such results is difficult owing to the complexities of the fingerprint data and the align-and-count algorithm. For instance, modeling the tendency for band size to be positively correlated within fingerprints and allowing for proportional error thresholds that vary by size seems prohibitive. More difficult is trying to model the underlying molecular biologic dynamics that produces fingerprint bands as a precursor to modeling (band) matching scores. This is because even a simple single base mutation that creates or destroys a restriction site can arbitrarily change the presence, absence, or size of a band. Far more complex changes affecting both the restriction sites and the *IS6110* elements, especially transpositions thereof, would need to be modeled. Because of these barriers, we proceed by a combination of empiricism, simulation and analogy, the analogy being to the

more tractable and developed world of sequence comparisons that is described next.

3.1 Sequence Comparison Scoring

The advent of rapid gene sequencing technology, the repository of resultant sequences in assorted databases and the subsequent "querying" of these databases for similarities to a new, target sequence has generated considerable statistical interest, especially with regard to assessing significance of the similarity scores so obtained. Recent overviews include Waterman and Vingron (1994) and Altschul and Gish (1996). Sequence data consists of strings of letters: for DNA sequences these come from the 4-letter alphabet (A, C, T, G) of nucleotides or bases, whereas protein sequences use the 20-letter alphabet of amino acids. Null characters are often included in the sequences to represent insertions or deletions. Given two sequences, algorithms exist (e.g., Smith and Waterman, 1981) for optimally aligning them. An optimal alignment consists of segments of the respective sequences that maximize a score function computed by totaling (i) matching scores for each pair of aligned letters and (ii) gap scores for each instance where a null character is aligned with a letter. We briefly summarize some relevant results on significance of optimal alignment scores, concentrating, for simplicity, on the case where gaps are not allowed.

Let $s(x, y)$ be the score for aligning letter x with letter y. For proteins, the scores may derive from structural or physicochemical properties. The nature of the optimal alignment depends critically on whether the expected score for a pair of randomly chosen letters is positive or negative. A positive expected score results in so-called global alignment: there is, on average, no penalty for accumulating additional mismatched letters so that the maximal score will correspond to virtually the entire shorter sequence. A negative expected score gives rise to the more interesting case of local alignment, on which we now focus. Expectations here, and distributional results to follow, are often based on assumptions that the letters are deterministic or iid. However, in efforts to achieve greater realism with regard to actual sequences, generalizations to Markov dependence have been made (Dembo and Karlin, 1991).

Let the maximal alignment score from comparing two random sequences of lengths m and n be

$$S_{m,n} = \max_{i,j,H} \sum_{h=0}^{H-1} s(x_{i+h}, y_{j+h}),$$

where the maximum is over all possible starting positions (i, j) and all possible alignment lengths H. Then a variety of analytic and empiric studies show that, for m and n sufficiently large, the distribution of $S_{m,n}$ is well represented by an extreme value distribution with cumulative distribution function

$$(1)\quad F_S(s) = \Pr\{S_{m,n} \le s\} = \exp(-\exp[-\lambda(s-u)]).$$

The parameters u and λ are termed the characteristic value and decay constant and correspond to location (mode) and inverse scale, respectively. Their estimation has been addressed using a variety of data and estimators. Among types of data considered are permuted or shuffled sequences, model-generated sequences and sequence databanks (these are often edited to exclude sequences known to be highly homologous). We discuss estimators below. Given parameter estimates, significance assessments for the comparison of either a pair of aligned sequences or, following some adjustment for multiple comparisons, a target sequence aligned to all sequences in a databank are readily obtained. For example, the $100(1-p)$ quantile S_{1-p} is estimated by

$$(2)\qquad S_{1-p} = \hat{u} - \hat{\lambda}^{-1} * \log(-\log(1-p)),$$

which is reasonably approximated by $S_{1-p} = \hat{u} - \hat{\lambda}^{-1} * \log(p)$ for $p \le 0.1$.

There are numerous software packages implementing a variety of alignment strategies and significance scoring assessments. Even though some refinements to significance assessment have been touted, the above extreme value theory still remains the workhorse and basis for most applications. A recent collection of papers, illustrating some advances, is provided by volume 266 of *Methods in Enzymology* (1996). Here, we are interested in applicability of these sequence-based ideas and results to fingerprint comparisons, and so we omit discussion of additional fine points.

3.2 Contrast and Analogy: Sequences and Fingerprints

A number of parallels between sequence and fingerprint matching scores suggest that extreme value distributional results from the former might pertain to the latter. Of course, such an assumption requires stringent checking. First, we outline the correspondences.

Alignment and scoring. At first blush, it appears that there are consequential differences between the settings with regard alignment and scoring, but this is not the case. Typically, for sequences, several hundred base pairs are compared, making appeal to large sample results reasonable. Fingerprints are aligned by incrementally sliding them past one another. As the score is the maximum over each of

these candidate alignments, and the number of increments is on the order of 100, we may again anticipate large sample results to obtain.

Sequence substitution scores are constructed to have negative expected value, thus favoring the emergence of local alignments. Further, properties of these scores feature in some derivations of extreme value asymptotics. Now, while it is the case that the expected value of the matching score scheme used for fingerprints (1 for bands within the prescribed tolerance, 0 otherwise) is positive, the distinction between local and global alignments for fingerprints is moot. The interest is in matching entire fingerprints, not portions thereof. Additionally, replacing the zero score with, say, -5 would not change which alignment scored maximally but would yield negative expected values.

Null model and data generation. With sequence data, the parameters of the extreme value distribution are often obtained from the set of scores from an artificial "null" databank. Using the resultant distribution for significance assessment is then deemed valid (Mott, 1992) because the query sequence will be unrelated to the vast majority of databank sequences. Generation of such databanks is either by way of (i) simulation or (ii) permutation. It is possible to incorporate varying degrees of local structure by using, for (i), prescribed probabilities not just for individual nucleotide or amino acid occurrence but also for joint events, and, for (ii), permutations within blocks of prescribed length. Alternatively, sets of scores from an appropriate (edited) referent databank can be used.

For fingerprints, artificial data generation is more complex. As permutation is not an option, we outline a simple simulation scheme below. However, it is apparent that this does not capture some essential features of fingerprints. Rather than trying to construct more intricate data-generation schemes, we revert to the use of referent databanks and resampling, with subsequent discussion (in Section 4) of editing issues.

To generate fingerprints we need a model. For simplicity, we base this on restriction site occurrence. The following is a highly simplified formulation:

1. Assume that the PvuII (CAGCTG) restriction sites (see Figure 1) occur at a constant rate outside the insertion sequence *IS6110*.
2. Assume a GC content of 65% (as observed), giving the probability of the restriction site sequence as $(0.35/2)^2(0.65/2)^4 = 0.000342$.
3. Assume that band sizes are distributed as a (truncated) exponential with rate parameter equal to the probability in item 2. The truncation arises because we cannot observe any band sizes less than approximately 900 bp.

It is very easy to efficiently generate band sizes according to the above model. We can then simulate fingerprints with varying numbers of bands. Figure 5 displays the empiric and model (both with and without truncation) band size distributions. While the model rate parameter (0.000342) agrees very closely with its empiric estimate (0.000344), there are obvious regions of lack of fit. These can in part be ascribed to preferred insertion sites of *IS6110*. These are regions of the genome partial to receiving *IS6110*. This results in the observed multiple modes. Also, second-order characteristics such as positive dependence between band sizes within fingerprints are lost. Remedying these and other deficiencies would require a highly contrived model. For these reasons, we emphasize use of resampling schemes as described in Section 4.

3.3 Extreme Value Parameter Estimation

Given a set of matching scores there are essentially three ways of estimating the parameters of the extreme value distribution: (i) moments, (ii) maximum likelihood and (iii) regression methods. The latter exploits hypothesized relationships between the parameters and the sample sizes, m, n, of the items (sequences, fingerprints) being compared (numbers of residues, number of bands, respectively). For example, in (1) one could assume that u grows linearly in $\log(mn)$ whereas λ is constant. Here, by virtue of the resampling schemes used, we have considerable data on conditional (on m, n) comparisons, and so focus on moment and likelihood estimation, thereby avoiding the regression assumptions. We then examine the behavior of the resultant estimates.

Method of moments estimators are given by

$$\tilde{\lambda} = \pi \Big/ \sqrt{6\,\mathrm{var}(S)}, \tag{3}$$

$$\tilde{u} = \bar{S} - \gamma/\tilde{\lambda}, \tag{4}$$

where γ is Euler's constant (0.577...) and $\bar{S}, \mathrm{var}(S)$ are the sample mean and variance, respectively. Because these estimators are one-to-one continuous functions of all (two) parameters they are consistent (e.g., Bickel and Doksum, 1977). Furthermore, they are obviously very easy to compute. Efficiency considerations are addressed in the Discussion (Section 5).

The density corresponding to (1) is $f_S(s) = \lambda \exp[-\lambda(s-u)] \exp(-\exp[-\lambda(s-u)])$. The attendant maximum likelihood estimates, $\hat{u}, \hat{\lambda}$, based on

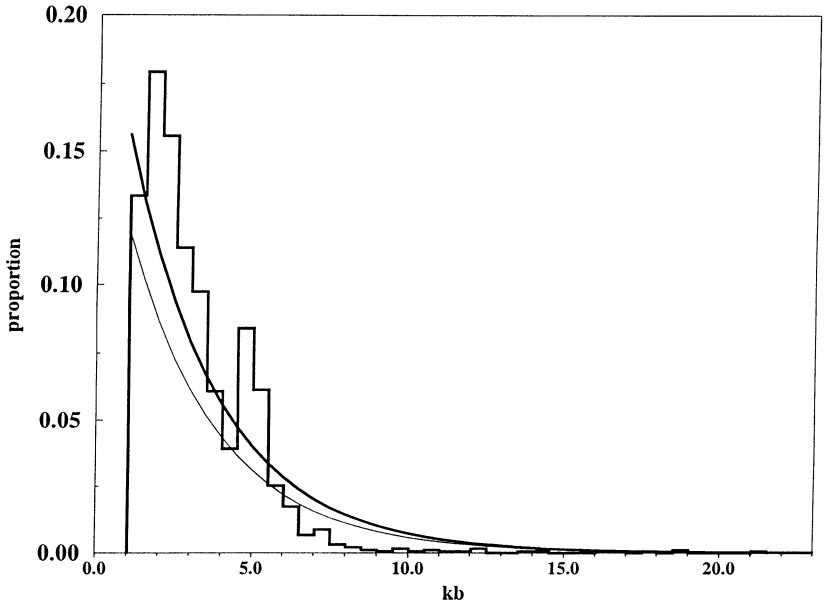

FIG. 5. *Histogram of band sizes for the San Francisco database*: *the heavy curve is the fitted truncated exponential model*; *the light curve is from an untruncated exponential model. For each model the rate parameter corresponds to the probability of observing the PvuII restriction site sequence assuming a GC content of* 65%.

a sample of scores $S_1, S_2, \ldots, S_k$ satisfy

$$\sum_{i=1}^{k} \exp\left(-\hat{\lambda}(S_i - \hat{u})\right) = k, \tag{5}$$

$$\sum_{i=1}^{k} (S_i - \hat{u}) * (1 - \exp\left(-\hat{\lambda}(S_i - \hat{u})\right)) = k\hat{\lambda}^{-1}. \tag{6}$$

Numerical solution is required. However, by adopting an approximation due to Kimball (1956), iteration may be avoided. The approximation works as follows. Substituting (5) in (6) yields

$$\begin{aligned} \hat{\lambda}^{-1} &= \bar{S} - k^{-1} \sum_{i=1}^{k} S_i \exp\left(-\hat{\lambda}(S_i - \hat{u})\right) \\ &= \bar{S} + k^{-1} \sum_{i=1}^{k} S_i \log(\hat{F}(S_i)), \end{aligned} \tag{7}$$

where $\hat{F}(S_i)$ is the estimated cdf which may be in turn be estimated from the data via the empiric distribution function using the order statistics $S_{(1)}, S_{(2)}, \ldots, S_{(k)}$. Approximating the expected value of $\hat{F}(S_{(i)})$ by the appropriate (continuity corrected) beta mean then gives the readily computable expression

$$\hat{\lambda}^{-1} = \bar{S} + k^{-1} \sum_{i=1}^{k} S_{(i)} \log\left(\frac{i - 1/2}{k + 1/2}\right). \tag{8}$$

Some comments on the maximum likelihood estimators are warranted. First, they are biased, but the biases in $\hat{u}$ and $\hat{\lambda}$ are compensatory with respect to quantile estimation (see (2)). Further, the biases are negligible when dealing with the large samples afforded by resampling. This is evident from the $O(k^{-1})$ bias correction formula given by Johnson and Kotz (1970). More consequential, however, is a lack of robustness reflected by the very large weight attached to small order statistics. An illustration of this effect is provided in the next section. This concern was presaged by Kimball (1956), before robustness considerations were in vogue.

3.4 Alternative Approaches to Standardization

Rather than relying solely on standardizations provided by extreme value distributions, we consider some alternate distributions. These include continuous (gamma) and discrete parametric densities. We also explore nonparametric methods. It is important to remain mindful of the goal of standardization: to readily provide a ranking of matching scores irrespective of the numbers of bands in the fingerprints being compared.

Gamma distribution. The two-parameter gamma distribution has density

$$f_S(s) = \frac{s^{\alpha-1} \exp(-s/\beta)}{\beta^{\alpha} \Gamma(\alpha)}$$

where $\alpha, \beta > 0$ and $\Gamma(\cdot)$ is the gamma function. Given a sample of scores $S_1, S_2, \ldots, S_k$ as above, let $Y = \log(\bar{S}) - k^{-1} \sum_{i=1}^{k} \log(S_i)$ (Y = log(arithmetic mean/geometric mean)). Then an approximate MLE

for α (Thom, 1968) is

$$\hat{\alpha} = \frac{1 + \sqrt{1 + 4/3Y}}{4Y}.$$

Bias corrections for $\hat{\alpha}$ exist, but these make little difference in the present context ($\hat{\alpha} \geq 10$). Given $\hat{\alpha}$, the MLE for β is $\hat{\beta} = \bar{S}/\hat{\alpha}$. Clearly, these estimates are easily computed.

The intended applications also require ready computation of quantiles and tail probabilities. This can be accomplished by adapting the Wilson–Hilferty transformation to the two-parameter gamma distribution:

$$S_{1-p} = \hat{\beta}\hat{\alpha}\left(\frac{\Phi^{-1}(1-p)}{3\sqrt{\hat{\alpha}}} + 1 - \frac{1}{9\hat{\alpha}}\right)^3, \tag{9}$$

where $\Phi(\cdot)$ is the standard normal cdf.

Given the similarity in their underlying shapes over a range of parameter values, we anticipate concordance between gamma and extreme value fits. Indeed, this transpires for moderate $m, n (\leq 10)$. But the two-parameter gamma tends to normality as $\alpha \to \infty$, which occurs here for $m, n = 15, 20$. Despite the gamma thus producing density estimates that are appreciably more symmetric than those provided by extreme value density estimates, there is nonetheless reasonable agreement between quantile estimates obtained with either family. Therefore, similar conclusions result from either parametric model.

Discrete distributions. Given the discrete support ($\{0, 1, \ldots, \min(m, n)\}$) of $S_{m,n}$, standardizing via discrete probability mass functions is indicated. Simplistic arguments can be proffered to justify use of binomial or hypergeometric families. However, perhaps due to this simplification (e.g., no accommodation for within-fingerprint dependence), generally poor tail fits were obtained. These were not improved by various generalizations. For example, attempting to account for differing band numbers in the fingerprints being compared by estimating the binomial sample size did not improve fit and yielded estimates close to $\min(m, n)$. Further, attempting to capture putative extra-binomial variation (arising from heterogeneity of matching probabilities) via, say, beta-binomial formulations failed due to $S_{m,n}$ being underdispersed for most ($\approx 80\%$) m, n combinations. Given the success of the continuous parametric families in meeting the standardization objectives, discrete alternatives are not pursued further.

Nonparametric approaches. Rather than prescribing parametric families, nonparametric approaches to standardization can be attempted. Figure 6 displays surfaces of $\Pr\{S_{m,n} = \min(m, n) - k\}$ against m and n for $k = 0, 1, 2, 3$ and $8 \leq m \leq n \leq 23$. Again, given the objectives of a unified (for all m, n and for k as given), accurate and readily computable standardization for the matching scores, both the irregular form of these graphs and results from using additive or local regression models (not shown) indicate that nonparametric approaches are less successful here than the simple parametric families employed and so will not be considered further.

4. RESULTS

Our results have a largely graphical flavor. After presenting a series of figures, we describe in the Discussion further implications of the findings. The basis for the various analyses are sets of align-and-count matching scores obtained either from simulation using the model from Section 3.2 or from resampling from the San Francisco database. This resampling was performed in a variety of ways: (i) individual band sizes were sampled and fingerprints with differing numbers of bands formed thereof; (ii) individual fingerprints were sampled; (iii) *distinct* individual fingerprints were sampled. In each instance sampling is with replacement. The idea behind the restriction ("editing") to distinct fingerprints in scheme (iii) is to exclude epidemiologically linked fingerprints. If the identical fingerprints so excluded arose from direct transmission of tuberculosis, their inclusion would distort the target "null" referent population. This distortion would conservatively bias significance assessments for matching a new fingerprint to the existing database. However, as our objectives are to assess (tail) adequacy of the extreme value distribution and examine attendant statistical issues, we focus primarily on scheme (ii) and note that results from schemes (ii) and (iii) were very similar. As discussed below, we believe this similarity derives from scheme (iii) only minimally accounting for epidemiologic linkage. Regardless of the scheme used, large datasets (sample sizes of 10^5) were generated. The align-and-count algorithm was then used to compute matching scores for all pairwise comparisons, distinguishing where applicable (schemes (ii) and (iii)) between intragel and intergel comparisons by using differing error tolerances.

Figure 7 displays line densities (histograms are not used, for clarity) for the number of matches corresponding to simulated data ("truncated exponential") and resampling schemes (i) and (ii) for the case of comparing two eight-banded fingerprints. Superimposed are extreme value fits (based on moment estimation) corresponding to scheme (ii) ("resampled fingerprints"). The agreement, where it matters

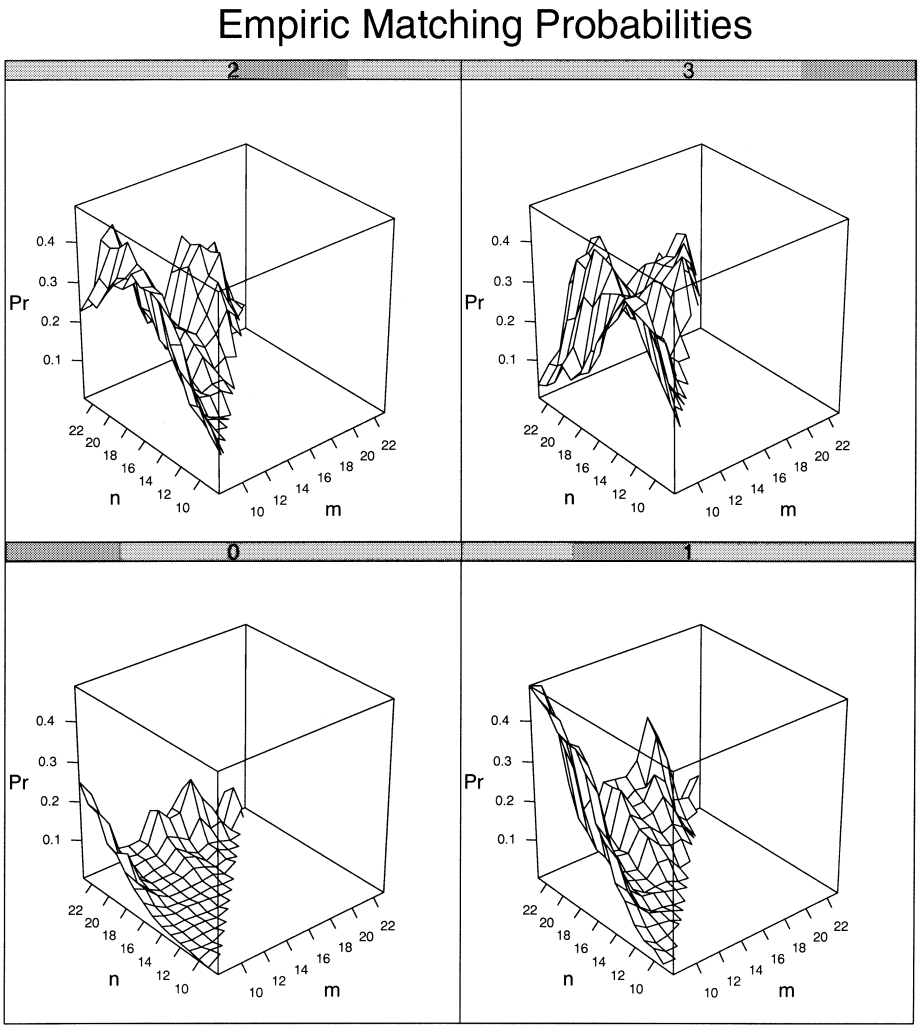

FIG. 6. *Empirical values of* $\Pr\{S_{m,n} = \min(m, n) - k\}$ *against* m *and* n *for* $k = 0, 1, 2, 3$ *and* $8 \le m \le n \le 23$: *the four panels correspond to the four values of* k.

in the right (upper) tail, is excellent. Such accurate approximation of the right tail also holds for the simulated and scheme (i) densities. The stochastic ordering of the densities is as anticipated. We observe more matching when resampling fingerprints because this is the only scheme that captures the positive dependence between bands within a fingerprint. That the simulated data are slightly stochastically smaller (less matching) than the resampled bands is simply due to the resampling approach usage of a finite set of band sizes.

We now focus solely on scheme (ii) and turn next to contrasting extreme value (likelihood and moment) and gamma fits. Revisiting the same eight- versus eight-band fingerprint comparison (Figure 8) we see near identical density estimates for the two extreme value estimators, attributable to the large sample sizes as noted in Section 3.3. The gamma distribution does worse in the right tail but better in the left tail.

We did not assume, however, that good fit or estimator concordance pertained irrespective of the numbers of bands in the fingerprint comparisons. Figure 9 shows (a) 10-versus 10-band, (b) 10-versus 15-band, (c) 15-versus 15-band and (d) 20-versus 20-band comparisons. In (a)–(c) the usefulness of the fitted extreme value densities in capturing the right tail is visually apparent. The gamma fits are comparable, with perhaps their (large sample) tendency toward symmetry producing poorer right tail fits [cf. (c), (d)]. But, in (d) (20 vs. 20) the tail approximation is considerably poorer. The differences between extreme value moment and likelihood estimates are also more pronounced. This behavior is

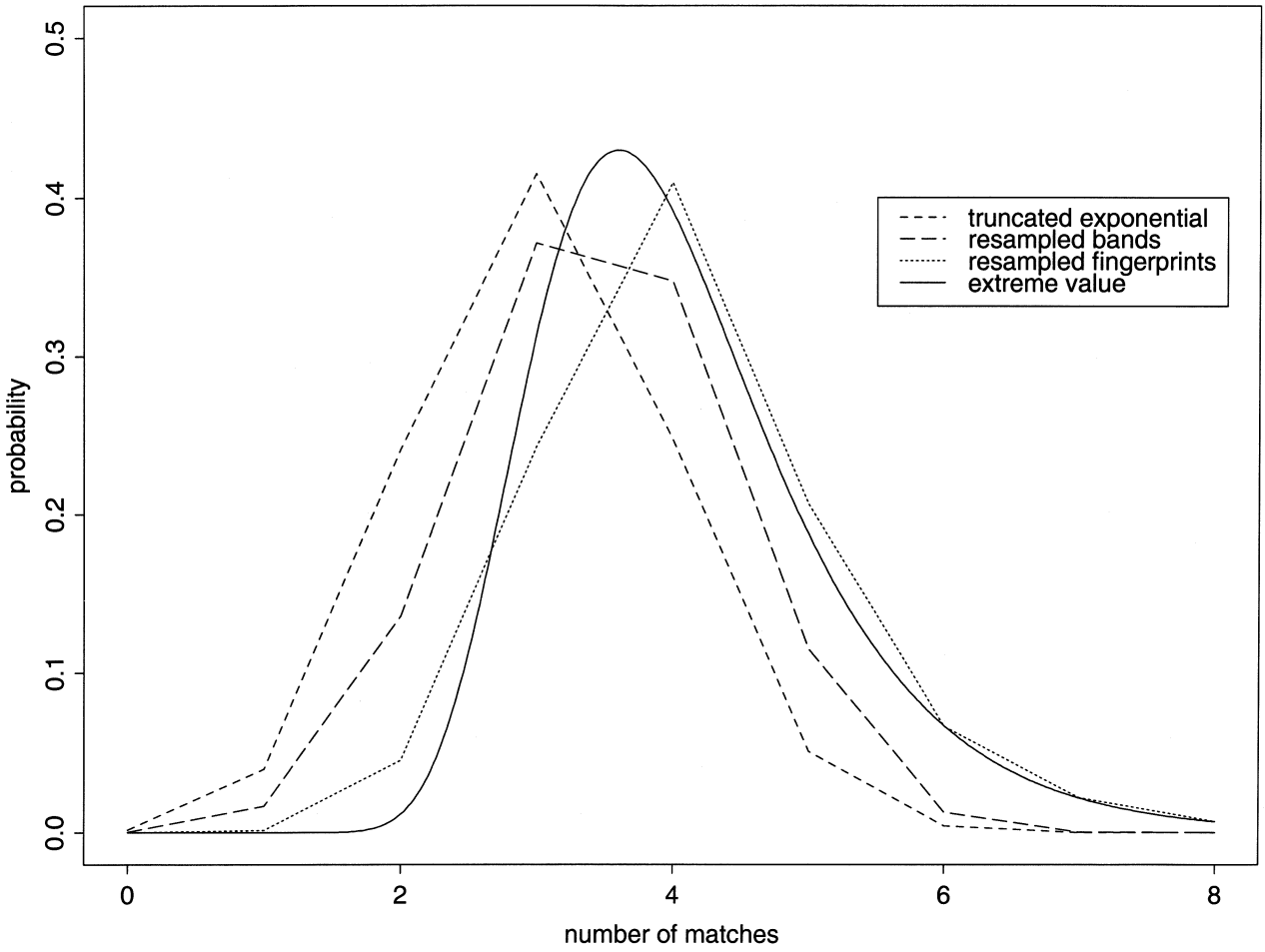

FIG. 7. *Densities of align-and-count matching scores for two eight-band fingerprints corresponding to differing fingerprint generation schemes (see text): the fitted extreme value density (solid) is based on resampled fingerprints.*

attributable to the left tail mass evident in the empiric density. As noted above, the MLE (8) attaches very large weight to small order statistics. The "excess" left tail mass itself results from there being few 20-band patterns (21 distinct) in the database so that the paucity of matches between select pairs emerges irrespective of how much resampling is performed. Further, as discussed below, relatedness among the few high band number patterns means that the mode and right tail are inflated, accentuating the left tail mass. The smoother, unimodal empiric densities obtained for the other comparisons reflect more diversity, for example, 80 distinct 10-band patterns.

To gain a better overview of extreme value and gamma fits to all possible (pairwise) fingerprint comparisons we next examine quantile estimates as a function of the number of bands constituting the comparison. Figure 10 displays 95% quantile estimates for both gamma and extreme value moment estimators according to (9) and (2), respectively. Again, there is little difference between the estimation methods. This also holds for extreme value quantile estimates based on maximum likelihood estimates. So, focusing on extreme value moment estimation, Figure 11 shows 95%, 99% and 99.9% quantiles versus the minimum of the number of bands in the two fingerprints being compared. Two interesting features emerge from this plot. The first concerns achievable significance levels for these sorts of comparisons. The maximal score obtainable is equal to the minimum of the number of bands, realized when all bands of the fingerprint with fewer bands match corresponding bands of the other fingerprint. Such scores are depicted by the 45° line. As is apparent, few comparisons achieve the 99% level, and none attain the 99.9% level. While we are not concerned with p-values

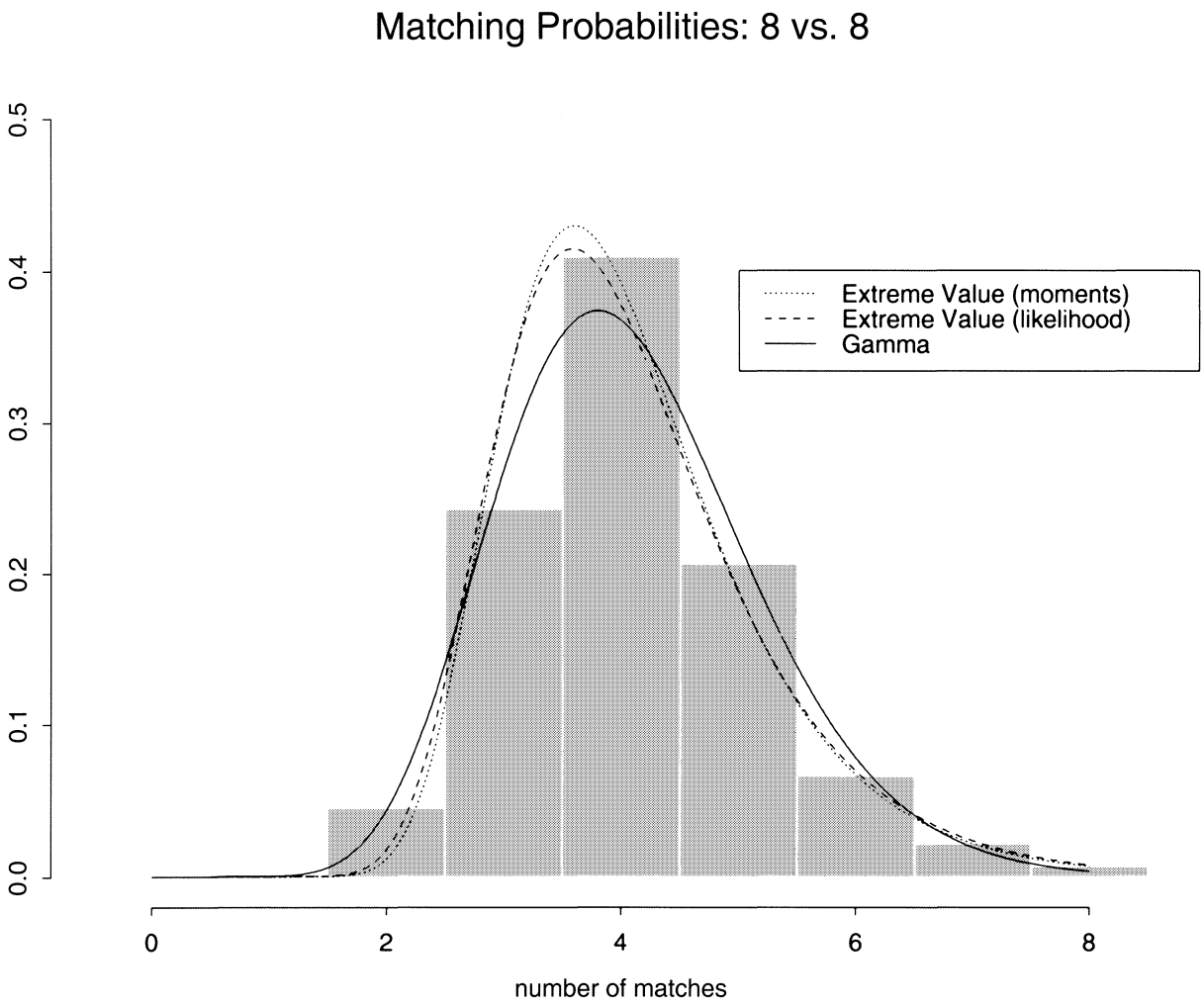

FIG. 8. *Extreme value and gamma densities for two eight-band fingerprints: the histogram depicts the empiric discrete density.*

per se, this is consequential from the perspective of invoking significance assessments when matching a fingerprint to a (large) database thereof. The need to accommodate multiple testing corrections, no matter how effected, will result in levels not reaching conventional goals, as is subsequently demonstrated.

Another finding pertains to the curvature of the smoothed quantile estimates, which is seen more clearly by restricting to comparisons of fingerprints with equal numbers of bands (not shown). The implications of this curvature are somewhat nonintuitive: more "significant" results are attainable when comparing two 15-band fingerprints than for two 20-band fingerprints. Initially, we attributed this result to precision losses deriving from band density within a fingerprint becoming too high: at some point matching within the prescribed tolerances becomes inevitable, leading to impaired resolution. However, investigation of precision phenomena using resampled bands revealed no such degradation (i.e., no curvature) out to 30 bands per fingerprint, a higher number than arises with tuberculosis. As the behavior of the two resampling schemes is the same in terms of precision, the explanation for the loss when using fingerprint resampling lies elsewhere. We have observed groups of visually similar fingerprints among high band number samples in the San Francisco data. That these are nonidentical limits the effectiveness of scheme (iii) above. Furthermore, high band number samples are more frequent among foreign born (relative to U.S.-born) subjects. The developing countries so represented have far higher rates of recent transmission. This all suggests epidemiologic linkage of these high band number strains

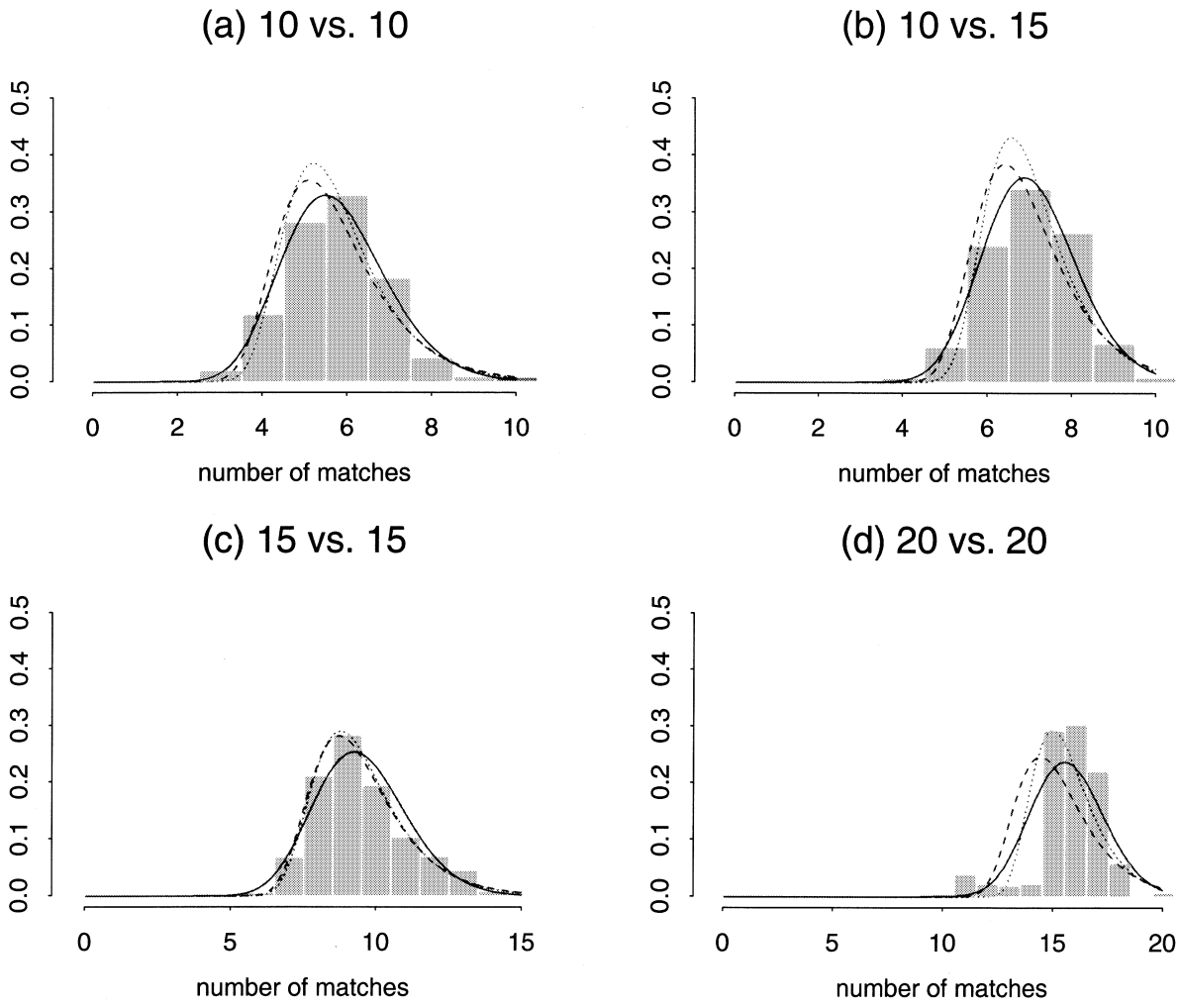

FIG. 9. *Extreme value and gamma densities for various band comparisons: the legend is as for Figure* 8.

which, in turn, results in the observed curvature by (i) inflating the estimate of the mode u (i.e., there is more matching), which in turn inflates the quantile $S_{1-\alpha}$ according to (2), and (ii) increasing variability (proportional to λ^{-1} by (3)) by reducing the number of distinct fingerprints. For comparison, there are 52 distinct 15-band fingerprints and, as noted, only 21 distinct 20-band fingerprints.

Figure 12 presents sample moments as functions of $\log(mn)$. As developed in the sequence setting, the simple theoretic forms of these relationships can be used as a basis for extreme value parameter estimation and as a diagnostic. In particular, under an extreme value distribution, the mean grows linearly in $\log(mn)$ while the standard deviation is constant. While agreement with the theoretic behavior is only passable, it is qualitatively comparable to that obtained in the sequence setting (Mott, 1992). Note that, as in Altschul and Erickson (1986), we have restricted attention to large (> 4.5) values of $\log(mn)$. These relationships permit estimation of global extreme value parameters by, say, nonlinear least squares. Here global means unconditional on the specific numbers of bands being compared. This offers the considerable benefit of avoiding simulation or resampling of the (numerous) band number specific comparisons.

We illustrate this and other points by comparing a collection of 99 fingerprints obtained in Orizaba, Mexico, with the San Francisco database. While the align-and-count algorithm allowed for rapid determination of all $99 \times 1{,}335 = 132{,}165$ matching scores, for the present purposes we focus on comparisons between the sole 15-band fingerprint from Orizaba with the 808 San Franciscan fingerprints having between 10 and 20 bands inclusive, yielding 808 matching scores $S_{15,n}, n = 10, \ldots, 20$. These were standardized using the extreme value

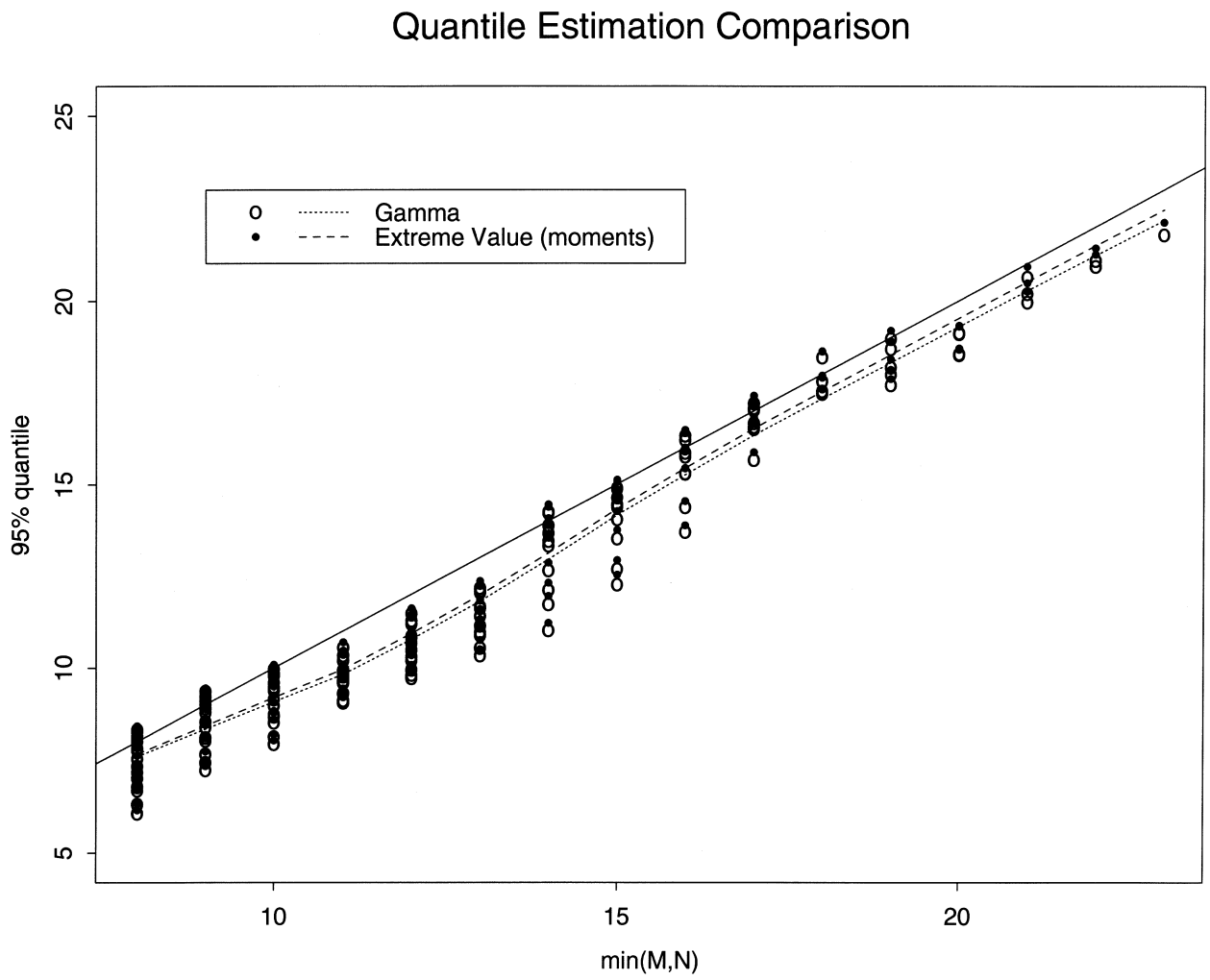

FIG. 10. *Comparison of gamma and extreme value (moments) 95% quantile estimates: the individual symbols correspond to distinct values of* $\max(m, n)$ *holding* $\min(m, n)$ *constant; The straight line represents the maximum achievable score which is the minimum of the two band numbers being compared.*

distribution. Moment estimation made recourse to fitting cubics [in $\log(mn)$] to the (resampled) means and standard deviations in Figure 12 and plugging into (3) and (4). The resulting matching probabilities are displayed in Figure 13. The (somewhat arbitrary) selection of the 15-band Orizaban fingerprint for illustrative purposes was based on the fact that it was the only fingerprint with a unique, yet moderate, number of bands.

Interesting findings include the following: (i) the "best" overall match corresponds to a 13-banded San Franciscan fingerprint with $S_{15,13} = 12$ and $\Pr\{S_{15,13} = 12\} = 0.009$ as compared to the best match with any 15-banded San Franciscan fingerprint for which $S_{15,15} = 13$ and $\Pr\{S_{15,15} = 13\} = 0.012$; (ii) correcting for the large number of comparisons, no matter how accomplished, would yield null results from a formal significance perspective—this remains so even if we were a priori to restrictively limit comparisons to the 52 distinct 15-banded San Franciscan fingerprints; and (iii) using the probabilities as a means for ranking as a prelude to further investigation reveals the following demographic information for the top three matches: a U.S. born white, a Honduran-born Hispanic and a Guatemalan-born Hispanic. Further follow-up would involve contact tracing.

5. DISCUSSION

The utility of genotyping tuberculosis, via DNA fingerprinting, has been widely demonstrated. Studies include investigation of outbreaks and multidrug-resistant strains (Edlin et al., 1992), transmission dynamics (Small et al., 1994) and lab cross-contamination (Small et al., 1993). Much of

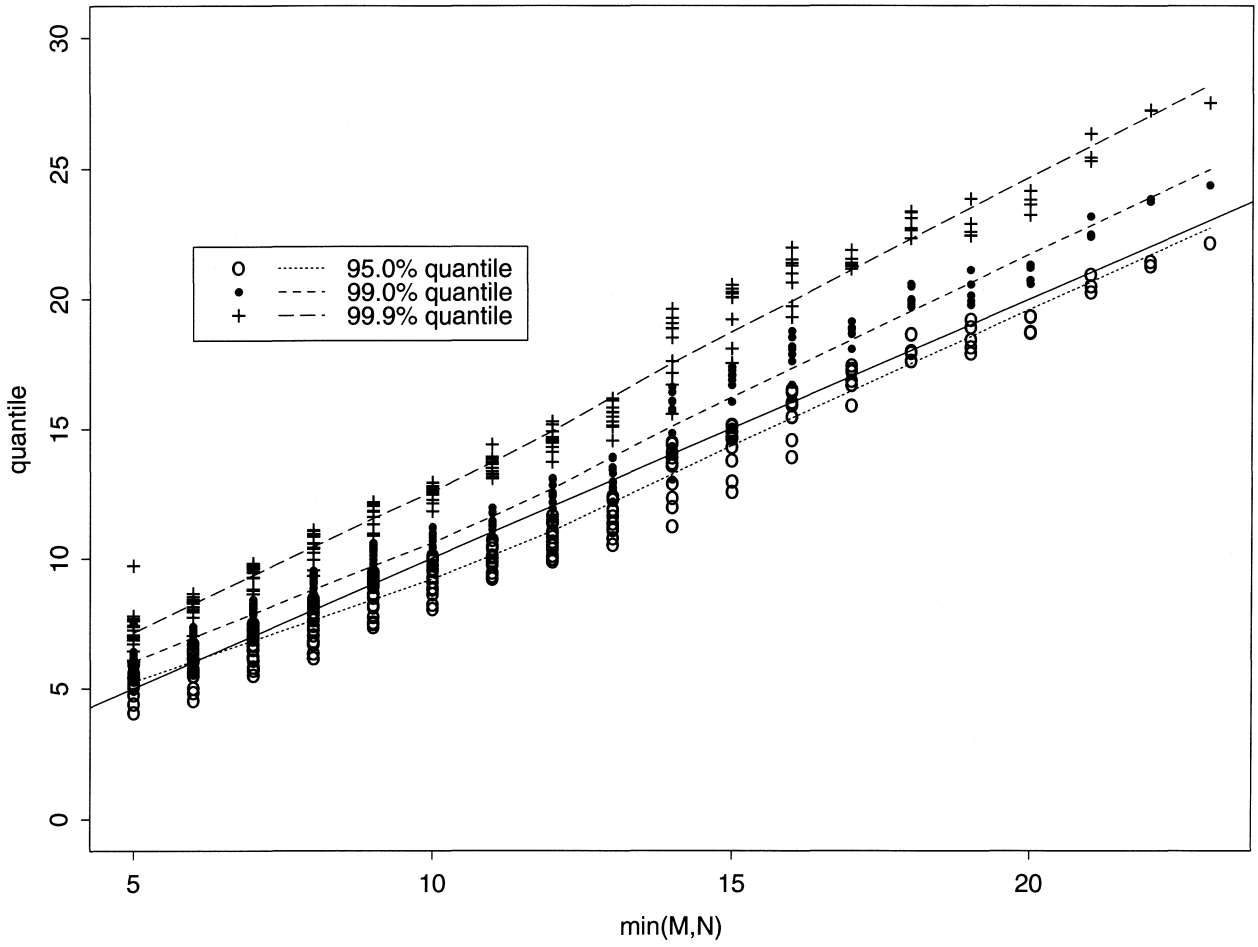

FIG. 11. *Moment estimators for select quantiles: straight line and individual symbols as in Figure* 10.

this work was done without recourse to the more careful alignment and assessment methods described here. So, we now consider the utility of such methods and consequent "added-value" in light of the above results.

We are not advocating the use of probabilities deriving from the fitted distributions for formal testing of fingerprint relatedness (more closely related organisms are those that have a shorter chain of transmission between them). Aside from all the assumptions and abuses posed thereby, as illustrated by the achievable significance plots and the Orizaba comparisons, such testing would only yield null results when multiplicity considerations were accommodated. Rather, the probabilities can be used as a means of ordering and investigating fingerprint relatedness. This usage corresponds to that advocated for sequence significance assessments (Waterman and Vingron, 1994).

As noted by a referee, a number of additional factors can contribute to the lack of achievable significance levels. These include (i) the number of distinct fingerprints constituting the database(s), (ii) epidemiologic linkage of these fingerprints, (iii) the appropriateness of referent densities and how they are fitted and (iv) the extent of information in the fingerprints themselves. While we have commented on some of these items a detailed examination of their relative importance is beyond the scope of this paper.

One potential usage is to treat the probability p as a similarity measure of fingerprint relatedness (or $1 - p$ as a distance). These can then be used as inputs to clustering or phylogeny algorithms. Phylogenies depicting strain relationships play a number of roles. First, the extent of clustering so delineated is used as a proxy for the percentage of tuberculosis attributable to recent infection, as opposed to latent reactivation (Small et al., 1994). This has

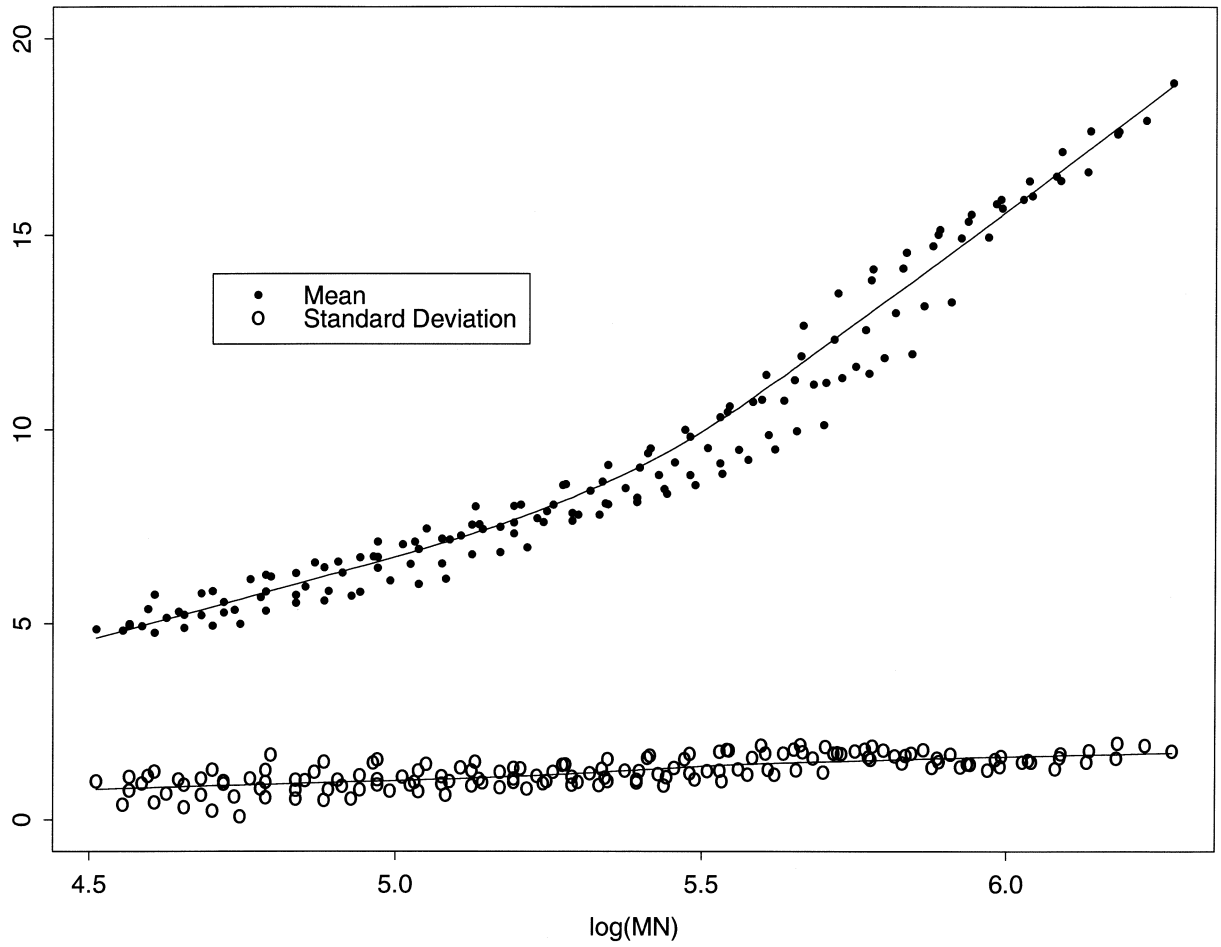

FIG. 12. *Sample means and standard deviations for align-and-count matching scores as a function of band numbers.*

direct public health implications. Presently, clustering is done in a very crude fashion. Improved distance measures ought to enhance the ability to delineate clusters, although the need for sensitivity analysis remains. We note that pursuing such sensitivity by bootstrapping the phylogeny (Newton, 1996; Efron, Halloran and Holmes, 1996) is problematic here because the bootstrap requires independence, or known dependence structure, among the attributes being resampled, neither of which pertains to fingerprint data. Second, fingerprint databases often contain, or are linked to, clinical and epidemiologic outcomes. It is desired to relate these outcomes to covariates. However, to the extent that strain affects outcome, the data are not independent but, rather, phylogenetically related. So, again, it is important to have good inputs for phylogeny estimation.

What extreme value or gamma fitting therefore provides is a basis for standardizing matching scores so they can be used in their totality. Concerns about the adequacy of this standardization for low or moderate scoring matches, based on the lack of fit evidenced in corresponding regions in some of the figures, are mitigated by the fact that the associated comparisons (i) are inconsequential with regard to clustering and (relatedly) (ii) are downweighted in phylogenetic regressions since they correspond to deep branches.

Interestingly, more bands do not necessarily provide more information as discussed above. There the issue of recent transmission resulting in an overrepresentation of related fingerprints—in this instance of high band number—was raised. These concerns relate to database editing and/or data generation. While differing approaches can dramatically impact resultant probabilities, the relative orderings are less affected. Determining an appropriate database on which to base parameter estimation and subsequent standardization depends heavily on context and, indeed, assumptions about whether relatedness derives from transmission or from molecular-level constraints.

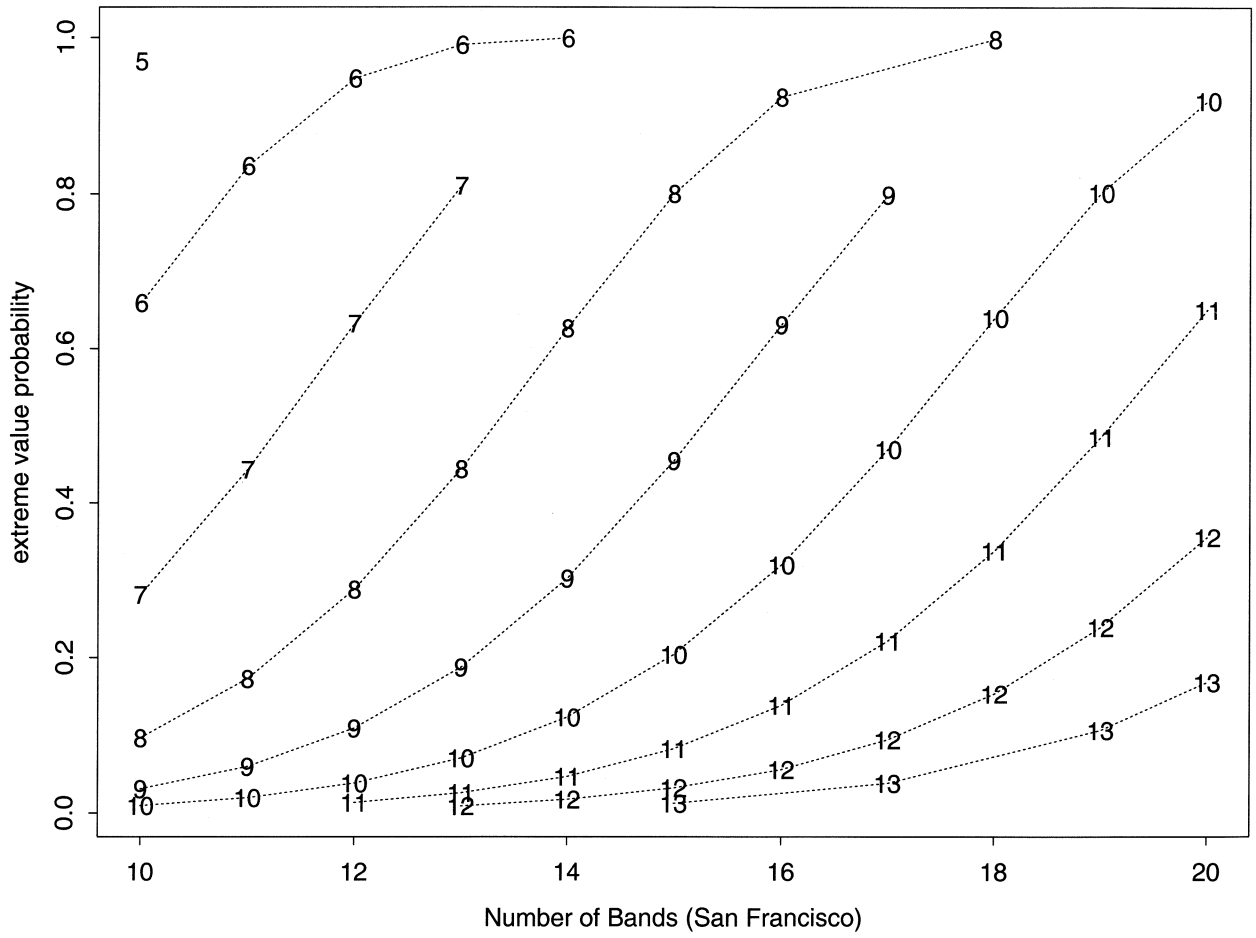

FIG. 13. *Extreme value probabilities for matching scores between the* 15*-band Orizaba fingerprint and* 10*- to* 20*-banded San Francisco fingerprints.*

Some comment on extreme value parameter estimation from the practical perspective is warranted. In most instances, very similar estimates were obtained from either moment or maximum likelihood estimation. This being the case, moment estimation would be preferred on the grounds of its computational ease. Beyond this, however, the robustness concerns cited in the context of Figure 9 make maximum likelihood less desirable. This is especially so in view of interest focusing on the right tail, a point noted by Kimball (1956). The objective of using resulting probabilities for the above standardization purposes makes efficiency considerations moot.

The utility of the align-and-count algorithm for comparing fingerprints hinges on the availability of fingerprint replicates. There is a general need to study sensitivity to the choice of strain used in generating replicates. Unfortunately, we are not positioned to do this as the only laboratory strain for which replicates were obtained was H37Rv as used. Qualities that the replications ought to possess include (i) covering the range of band sizes that will be measured in experimental samples, (ii) multiple (ideally random) placements in lanes across the electrophoresis gel and (iii) minimized opportunities for evolution. H37Rv possesses properties (i) and (iii) and is satisfactory with regard to (ii).

The approach taken here is shamelessly empiric. That is, both scoring via the align-and-count algorithm and subsequent standardization via extreme value or gamma distributions were motivated by empiric as opposed to model-based considerations. However, this approach was not pursued on grounds of expediency. Rather, devising an appropriate molecular biologic–statistical model appears prohibitive given existing data due to the following concerns. As mentioned, the fact that a single base mutation can arbitrarily change the presence, absence, or size of a band and more complex changes can impact both the restriction sites and *IS6110* el-

ements makes modeling problematic. This contrasts with fingerprint data obtained solely from restriction enzyme cutting (Nei and Li, 1979) although even there the modeling is simplistic. Further, at some point, an empiric approach will be necessary to accommodate the various sources of measurement error including those attributable to alignment. Attempting to model attendant within-fingerprint dependencies is also problematic as exemplified by (i) the unstructured pattern of H37Rv correlations (not shown) and (ii) the need to condition on band numbers and sizes. One potential (future) source from which we hope to obtain data allowing modeling of band pattern dynamics is the repeated fingerprinting of individuals with long term, active tuberculosis. This data may permit, if sufficiently rich, probability and rate assessments for band pattern changes.

ACKNOWLEDGMENTS

Support for this work was provided by NIH Grants AI40906 and AI34238. The authors thank two anonymous referees and the Editor for numerous suggestions that greatly improved both content and presentation.

REFERENCES

ALTSCHUL, S. F. and ERICKSON, B. W. (1986). A nonlinear measure of subalignment similarity and its significance levels. *Bull. Math. Biol.* **48** 617–632.

ALTSCHUL, S. F. and GISH, W. (1996). Local alignment statistics. *Methods in Enzymology* **266** 460–480.

BICKEL, P. and DOKSUM, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco.

CENTERS FOR DISEASE CONTROL AND PREVENTION. (1994). *Addressing Emerging Infectious Disease Threats: A Prevention Strategy for the United States*. U.S. Department of Health and Human Services, Public Health Service, Atlanta, GA.

DEMBO, A. and KARLIN S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. *Ann. Probab.* **19** 1737–1755.

EDLIN, B. R., TOKARS, J. I., GRIECO, M. H., CRAWFORD, J. T., WILLIAMS, J., SORDILLO, E. M., ONG, K. R., KILBURN, J. O., DOOLEY, S. W. and HOLMBERG, S. D. (1992). An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *New England J. Medicine* **326** 1514–1521.

EFRON, B., HALLORAN, E. and HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Nat. Acad. Sci. U.S.A.* **93** 7085–7090.

ERIKSEN, B. and SVENSMARK, O. (1993). DNA profiling of strains in criminal cases: analysis of measurement errors and band shift. *Forensic Sci. Internat.* **61** 21–34.

INSTITUTE OF MEDICINE. (1992). *Emerging Infections: Microbial Threats to Health in the United States*. National Academy Press, Washington, DC.

JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions* **1**. Wiley, New York.

KIMBALL, B. F. (1956). The bias in certain estimates of the parameters of the extreme value distribution. *Ann. Math. Statist.* **27** 758–767.

MOTT, R. (1992). Maximum likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bull. Math. Biol.* **54** 59–75.

NEI, M. and LI, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Nat. Acad. Sci. U.S.A.* **76** 5269–5273.

NEWTON, M. A. (1996). Bootstrapping phylogenies: large deviations and dispersion effects. *Biometrika* **83** 315–328.

ROEDER, K. (1994). DNA fingerprinting: a review of the controversy. *Statist. Sci.* **9** 222–278.

SALAMON, H., SEGAL, M. R. and SMALL, P. M. (1998). Automated comparison and clustering of bacterial DNA fragment-based genotypes. *Emerging Infectious Disease* **4** 159–168.

SMALL, P. M. (1995). Towards an understanding of the global migration of tuberculosis. *J. Infectious Disease* **171** 1593–1594.

SMALL, P. M., HOPEWELL, P. C., SINGH, S. P., PAZ, A., PARSONNET, J., RUSTON, D. C., SCHECTER, G. F., DALEY, C. L. and SCHOOLNIK, G. K. (1994). The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. *N. England J. Medicine.* **330** 1703–1709.

SMALL, P. M., MCCLENNY, N. B., SINGH, S. P., SCHOOLNIK, G. K., TOMPKINS, L. S. and MICKELSEN, P. A. (1993). Molecular strain typing of *Mycobacterium tuberculosis* to confirm cross-contamination in the AFB laboratory and modification of procedures to minimize occurrence of false positive cultures. *J. Clinical Microbiol.* **31** 1677–1682.

SMITH, T. F. and WATERMAN, M. S. (1981). Identification of common molecular subsequences. *J. Molecular Biol.* **147** 195–197.

SUDBURY, A. W., MARINOPOULOS, J. and GUNN, P. (1993). Assessing the evidential value of DNA profiles matching without the assumption of independent loci. *J. Forensic Sci. Soc.* **33** 73–82.

THOM, H. C. S. (1968). *Direct and Inverse Tables of the Gamma Distribution*. Environmental Data Service, Silver Spring, MD.

VAN EMBDEN, J. D. A., CAVE, M. D., CRAWFORD, J. T., DALE, J. W., EISENACH, K. D., GICQUEL, B., HERMANS, P., MARTIN, C., MCADAM, R. and SHINNICK, T. M. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J. Clinical Microbiol.* **31** 406–409.

WATERMAN, M. S. and VINGRON, M. (1994). Sequence comparison significance and Poisson approximation. *Statist. Sci.* **9** 367–381.

WOELLFER, G. B., BRADFORD, W. Z., PAZ, A. and SMALL, P. M. (1995). A computer assisted molecular epidemiologic approach for confronting the re-emergence of tuberculosis. *Amer. J. Medical Sci.* **311** 17–22.