Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infection, Genetics and Evolution 5 (2005) 35–43
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information

Mark M. Tanaka [a], Andrew R. Francis [b,*]

[a] *School of Biotechnology and Biomolecular Sciences, University of New South Wales, NSW 2052, Australia*
[b] *School of Quantitative Methods, and Mathematical Sciences, University of Western Sydney, Locked Bag, 1797, Penrith South DC, NSW 1797, Australia*

Received 25 February 2004; received in revised form 4 June 2004; accepted 4 June 2004
Available online 22 July 2004

Abstract

Genotypic data from pathogenic isolates are often used to measure the extent of infectious disease transmission. These methods include phylogenetic reconstruction and the evaluation of clustering indices. The first aim of this paper is to critique current methods used to analyse genotypic data from molecular epidemiological studies of tuberculosis. In particular, by not accounting for the mutation rate of markers, errors arise in making inferences about outbreaks based on genotypic information. The second aim is to suggest a new way to represent genotypic data visually, involving graphs and trees. We also discuss some interpretations and modifications of existing indices. Although our focus is tuberculosis, the methods we discuss are generally applicable to any directly transmissible clonal pathogen.

Keywords: Cluster analysis; Genetic markers; Molecular epidemiology; Mutation; Disease outbreaks; *Mycobacterium tuberculosis*; Graph theory; Data visualisation

1. Introduction

Molecular tools have enhanced efforts by epidemiologists to characterise the spread of infectious diseases such as tuberculosis. The genotyping of isolates has allowed the investigation of several important questions. For example, epidemiological links inferred from contact tracing can be supported, potential risk factors can be studied, and conclusions about the population-level state of epidemics can be drawn (Foxman and Riley, 2001; Seidler et al., 2004).

As molecular epidemiological data accumulate, we are presented with the following challenge: how do we quantify the severity of an epidemic using the genotypic data? In the case of tuberculosis (TB), this question is closely related to that of estimating the proportion of TB cases due to recent transmission rather than to the reactivation of latent infections. Current methodologies for analysing genotypic data from pathogenic isolates involve first grouping isolates into clusters of identical genotypes, then computing a phylogeny or an index of clustering based on the sizes of these clusters (to be discussed later). One of the applications of these approaches is to identify risk factors linked to "clusteredness" and therefore elevated disease transmission.

As genotyping technologies develop, the resolution of the genetic information that can be extracted from pathogenic isolates increases. For example if the number of genetic sites considered increases, existing clusters will become partitioned into smaller clusters, and the overall mutation rate for these sites will increase. This change may have dramatic effects on clustering indices. A number of studies have suggested the use of additional markers to increase the resolution of currently available technologies. For example, because low IS*6110* copy-number strains of *M. tuberculosis* yield little information, these studies have discussed the use of other markers such as spoligotyping to refine the data (Bauer et al., 1999; Soini et al., 2001; Rhee et al., 2000).

A further issue is that of mutation events occurring within the time frame of an epidemiological investigation. A high mutation rate will tend to reduce the sizes of the clusters in a sample of genotypes from an outbreak. Combining the information from several markers raises the overall mutation rate across those markers. In fact, it is known that even alone IS*6110* evolves at a rate high enough to produce mutation events on short time scales (Yeh et al., 1998; Niemann et al., 1999; Tanaka and Rosenberg, 2001; Rosenberg et al., 2003; van der Spuy et al., 2003).

* Corresponding author. Tel.: +61-2-9852-4152; fax: +61-2-9852-4103/4185.
E-mail address: a.francis@uws.edu.au (A.R. Francis).

1567-1348/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2004.06.001

This article investigates methods for interpreting genotypic data from clonal pathogens such as tuberculosis. We are concerned with extracting information from molecular data, rather than estimating disease incidence from data obtained through, for example, tuberculin skin test (TST) surveys. In Section 2 we describe some existing approaches to analysing the population structure of *M. tuberculosis* based on pathogen genotype data sets. We also offer some cautionary comments about these techniques in Section 2.3 based on considerations of the mutation process. Section 3 introduces a way to visualise molecular epidemiological data, which also provides a framework for interpreting some of the current clustering indices. The scope of the methods discussed here is limited to clonal (asexual) pathogens that are transmitted directly between hosts. Organisms that undergo genetic exchange can be included if recombination can be detected as discrete events and the recombination rate is known. That is, for the methods discussed here to be applicable, it must be possible to treat recombination itself as a form of mutation.

2. Methods of quantifying outbreaks of clonal pathogens using genotypic data

2.1. Recent transmission indices from molecular epidemiological studies of tuberculosis

We will discuss two related indices of outbreak severity used in the molecular epidemiology of tuberculosis. We will refer to both of these as recent transmission indices because they are intended to reflect the extent of recent transmission of tuberculosis. The first index is $\mathrm{RTI}_{n-1} = (n_c - c)/n$ (Small et al., 1994), where n is the total number of cases in the sample, c is the number of genotypes represented by at least two cases, and n_c is the total number of cases in cluster of size two or greater. Here, the numerator counts the number of transmissions reflected by the data set (c is the number of "source" cases: one per cluster). Similarly, a second index $\mathrm{RTI}_n = n_c/n$ has been used (Alland et al., 1994). These have been referred to respectively as the "$n-1$ method" (RTI_{n-1}) and the "n method" (RTI_n) (Glynn et al., 1999; Murray, 2002). Both indices capture the intuitive idea that the more genetically homogeneous a data set is, the more severe the extent of transmission.

We now make some observations about these indices. First, define n_i to be the number of cases in the ith genotype, and let g be the number of different genotypes in the sample. The total number of cases is $\sum_{i=1}^{g} n_i = n$. The fact that $n_c - c = \sum_{i=1}^{g}(n_i - 1)$ motivates the term "$n-1$ method". Defining u to be the number of genotypes that are unique in the sample ("singletons"), note that $n_c - c = (n_c + u) - (c + u) = n - g$, so that $\mathrm{RTI}_{n-1} = (n-g)/n = 1 - (g/n)$. This shows that RTI_{n-1} is dependent only on the sample size and the number of genotypes. We remark that strictly speaking, RTI_{n-1} cannot equal 1, since there is always at least one genotype in the sample ($g \geq 1$). To use this index of recent transmission strictly as a proportion, we suggest the following minor adjustment, so that the index has a maximum of 1, and a minimum of 0.

$$\mathrm{RTI}^{*}_{n-1} = \frac{n-g}{n-1} = 1 - \frac{g-1}{n-1}.$$

Turning now to RTI_n, we note that since $n_c = n - u$, it can be rewritten as follows: $\mathrm{RTI}_n = (n-u)/n = 1 - (u/n)$. RTI_n is therefore dependent only on the sample size and the number of singletons.

2.2. Diversity measures from ecological studies

Because clustering can be viewed as the opposite of diversity, we mention measures of diversity often used in ecological research. Simpson's index is the probability that any two individuals (isolates) chosen at random from a data set are from the same species (genotype):

$$S = \sum_{i=1}^{g} \frac{n_i(n_i - 1)}{n(n-1)},$$

where n_i is the number of individuals in species i (genotype i) in the sample. This index is a measure of clustering, with range [0,1]. Simpson's diversity index ($1 - S$ and its variants such as sampling with replacement) are also used in population genetics, where they are called heterozygosity and gene diversity and are used to study global patterns of variation (e.g. Selander and Levin, 1980).

Similarly, the Shannon-Weaver index, from information theory, is used to describe the diversity of ecological communities. This index is given by $H = -\sum_{i=1}^{g}(n_i/n)\ln(n_i/n)$. Because the greatest H possible is $\ln(n)$, a clustering index based on this quantity (normalising the range to be [0,1]) would be

$$C_H = 1 - \frac{H}{\ln(n)} = \sum_{i=1}^{g} \frac{n_i \ln(n_i)}{n \ln(n)}.$$

The index $1 - S$ has been used in molecular epidemiology to measure how well a genetic marker discriminates strains (Hunter and Gaston, 1988; Dale et al., 2003). We remark that both S and C_H can be used to measure the genetic homogeneity of a set of pathogenic genotypes (though we could not find examples in the literature).

One noteworthy difference between the RTI indices and these indices is that the latter use all of the cluster sizes. At this stage it is not clear whether this is of critical import for making inferences about the epidemic.

2.3. Genetic heterogeneity and mutation rate

Because the mutation rate of the marker affects the configuration of clusters in a sample, conclusions about the epidemic based on clusters may be prone to error unless mutations are considered. A highly heterogeneous data set may

Table 1
Inferences about an outbreak based on the homogeneity of data

Mutation Rate	Data	
	Homogeneous	Heterogeneous
High	Severe outbreak	Ambiguous
Low	Ambiguous	Mild outbreak

lead to an underestimate of the speed of the outbreak if the mutation rate is high, since the reason for the observed diversity may be rapid mutation rather than slow transmission. Conversely, while transmission may appear to be fast because of a genotypically homogeneous data set, the observed clustering may simply indicate a slow mutation process.

Furthermore, ignoring mutation may lead to false outcomes when comparing different data sets. For example, a genetically heterogeneous data associated with a high mutation rate may conceivably represent a more severe outbreak than a homogeneous data set associated with a low mutation rate, as summarised in Table 1. Existing clustering indices are unable to detect this possibility.

One consequence of this problem is that direct comparison of data sets generated using different markers is not possible. This is partially recognised in molecular epidemiology when it is noted that some markers have greater discriminatory power than others (Tenover et al., 1995; Kremer et al., 1999; Coenye et al., 2002). As new markers are developed and replace older techniques the information gathered through the old methods will be made redundant (Foxman and Riley, 2001). It would be desirable to retain such information, for example by incorporating mutation rates into clustering indices.

A basic mutation model and the graph-theoretic organisation of the data (as described later, in Section 3.2 and Appendices A and B) can be used to derive a simple alternative to the RTI which we call the transmission mutation index (TMI):

$$\mathrm{TMI} = \tilde{\mu}\frac{(n - g + \nu_1)}{\nu_1}, \quad (1)$$

where $\tilde{\mu}$ is an independent estimate of the mutation rate of the marker, and ν_1 is the number of single-step mutation events inferred from the data. A more formal definition of ν_1 is given in Appendix B.

Although the TMI uses extra pieces of information (ν_1 and $\tilde{\mu}$) and thus appears to moderate the sensitivity to μ, it needs to be further examined. Possible limitations are that it may be over-sensitive to small values of ν_1, and the mutation rate of a genetic marker is not always known with precision.

3. Approaches to visualising genotypic data from clonal pathogens

3.1. Current graphical techniques

The grouping of isolates into clusters of identical genotypes—as discussed above—does not make much use of the genetic relationships among the isolates. At the other extreme, there have been efforts to organise all genotypes present in a sample according to genetic relationships, i.e., by constructing phylogenies (e.g. Duchene et al., 2004). A popular method for analysing DNA fingerprints is the use of the Dice coefficient for assigning distances between DNA fingerprints, in conjunction with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbour-Joining for grouping genotypes (e.g. Toungoussova et al., 2002). Once a tree or dendrogram is drawn it can be eyeballed to locate similar cases that "cluster" together. Although these trees are not necessarily intended to represent evolutionary relationships, they are highly suggestive of evolutionary trees.

When there is insufficient information in the genetic marker, the use of phylogenies is problematic. There are large uncertainties involved in the estimation of distance. Furthermore, it is not clear how functions such as Dice are related to time, but the relationships are unlikely to be linear. That is, the distance measure may not adequately reflect evolutionary divergence so that time is distorted in the inferred trees. Thus, in this context, dendrograms attempt to extract more information from the marker than it carries. They also make little use of information about genotypically identical isolates.

A compromise between only using clusters of identical genotypes and drawing phylogenies is to construct graphs with genotypes at internal nodes, edges reflecting direct evolutionary relationships, and cluster sizes indicated at the nodes. In the context of infectious disease outbreaks, these latter developments promise improved insight into recent epidemiological history. For instance, the algorithms BURST and eBURST have been applied to multilocus sequence typing (MLST) data from a variety of bacterial species (Feil et al., 2004). A related idea was used by Zhu et al. (2001) for *Neisseria meningitidis*. Similar graphical visualisations called Minimum Spanning Networks have been used to study population structure in eukaryotic organisms (Excoffier et al., 1992; Excoffier and Smouse, 1994).

In the study of Duchene et al. (2004) two further techniques, Median-Joining Networks and Cladistic Nested Analysis, were deployed and compared in the analysis of spoligotypes from four Caribbean islands. In Section 3.3 we will re-analyse their data using a new approach outlined in Section 3.2.

3.2. Cluster-graphs and transmission trees

In this section we will describe a new graph-theoretic construct for representing and analysing clonal pathogenic genotypes.

Given a data set of isolates we first form clusters of g distinct genotypes. Consider now a graph in which each vertex

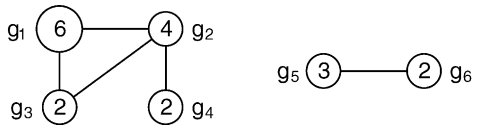

Fig. 1. A cluster-graph representing a sample with five distinct genotypes. Each node represents a cluster of genotypically identical isolates, and the sizes of clusters are indicated inside the circles. There are two connected components in this example.

represents a cluster. Connect all pairs of vertices whose genotypes are separated by a single mutation step. This is what we will call a *cluster-graph*. For example, suppose there are six distinct genotypes, $g_1 \cdots g_6$, respectively with 6, 4, 2, 2, 3, 2 isolates of each type. Suppose there are single mutation steps between the following pairs of genotypes: (g_1, g_2), (g_1, g_3), (g_2, g_3), (g_2, g_4), (g_5, g_6), but not between any other pairs. Then the cluster-graph would be as in Fig. 1. Note that a cluster-graph will not necessarily be connected.

We now define a second kind of graph based on cluster-graphs that can be used to interpret measures of clustering. A *transmission tree* spans all the isolates in a connected component of the cluster-graph. The number of edges in such a tree equals the number of vertices in the connected component minus 1. Each edge represents a possible transmission event between the vertices at either end of the edge. The entire data set can be viewed as a transmission forest.

In Fig. 2 we show an example of a possible transmission forest based on the cluster-graph of Fig. 1. The biological basis for this construct is discussed in Appendix A. The aim here is not to infer the actual history behind the data set as a transmission forest; rather, it is to use the concept to summarise and interpret the data.

3.3. Analysis of spoligotype data: an example

In this section we produce a cluster-graph representing data from Duchene et al. (2004) to demonstrate the methods discussed in Section 3.2. These data consist of 321 isolates representing 47 distinct alleles from four Caribbean islands. Spoligotyping in general produces data which are richer in close genetic relationships than for example IS*6110*. This aspect of spoligotyping means data thus obtained reveal

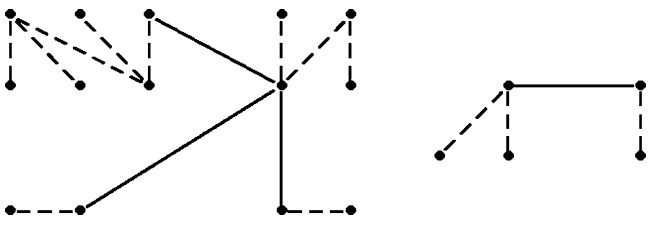

Fig. 2. One possible transmission tree based on the cluster-graph of Fig. 1. Dashed lines represent transmissions within clusters, and solid lines represent transmissions between clusters, involving a single mutation.

more about evolutionary connections between genotypes, and thus makes the cluster-graphs more informative. The Duchene et al. (2004) data set is suitable for our purposes as it represents an outbreak from a particular geographical region.

The steps taken in the construction of the cluster-graph were as follows. (1) Identify pairs of genotypes that can be connected by a single mutation event. In the case of spoligotypes, this means a single deletion event. (2) Names of genotypes and other relevant information label the vertices. Circles at vertices may be related to the size of the cluster. (3) Edges are drawn whenever two spoligotypes can be connected by a single deletion event. Where the direction of mutation is known, as is the case for spoligotypes, these edges can be drawn as arrows. Note, the final graph will not necessarily be a tree or a forest. Cluster-graphs are similar to the BURST representation of genotypes, except that here we do not attempt to resolve graphs into trees, and the direction of mutation is based on the presumed deletion process in spoligotypes.

Fig. 3 shows the cluster-graph for the Duchene et al. (2004) data set. Note that we did not attempt to resolve evolutionary histories into trees. There may be little advantage in making these additional inferences, and in fact some information may be lost. For example, where the graph indicates two possible paths, it may not be feasible to determine which one evolution took. Indeed, it may even be that homoplasy (independent mutation events producing the same genotype) occurred.

Observe that the use of the cluster-graph reveals the ancestral genotype to be the spoligotype labelled **53**. In agreement with the original reference, the three major clades identified by Duchene et al. (2004) are clearly visible in our cluster-graph (Fig. 3). These are the Haarlem superfamily (spoligotypes connected to types **50** and **47**, below type **53**), the LAM superfamily (spoligotypes connected to type **42** above type **53**), and the X superfamily (spoligotypes connected to type **119** to the right of type **53**). Finally, a fourth miscellaneous set of genotypes includes the putative ancestral type **53**.

Incidentally, applying the indices previously discussed to this data set gives the following values: $\mathrm{RTI}_n = 0.935$, $\mathrm{RTI}_{n-1} = 0.854$, $S = 0.066$, $C_H = 0.464$, and TMI = 0.31. The RTI values here are very high compared to the same statistic computed on IS*6110* data. For example, the data from Small et al. (1994) gives $\mathrm{RTI}_{n-1} = 0.311$. This difference is presumably due to the very different mutation rates of the two markers involved, and not to the levels of recent transmission. Support for this presumption can be found in the observation that IS*6110*-typing discriminates isolates more finely than spoligotyping (Kremer et al., 1999). This outcome demonstrates the problem concerning the mutation rate of markers described in Section 2.3. Note also that to compute TMI requires knowledge of the mutation rate, an accurate estimate of which is not currently available. In order to obtain the TMI value above,

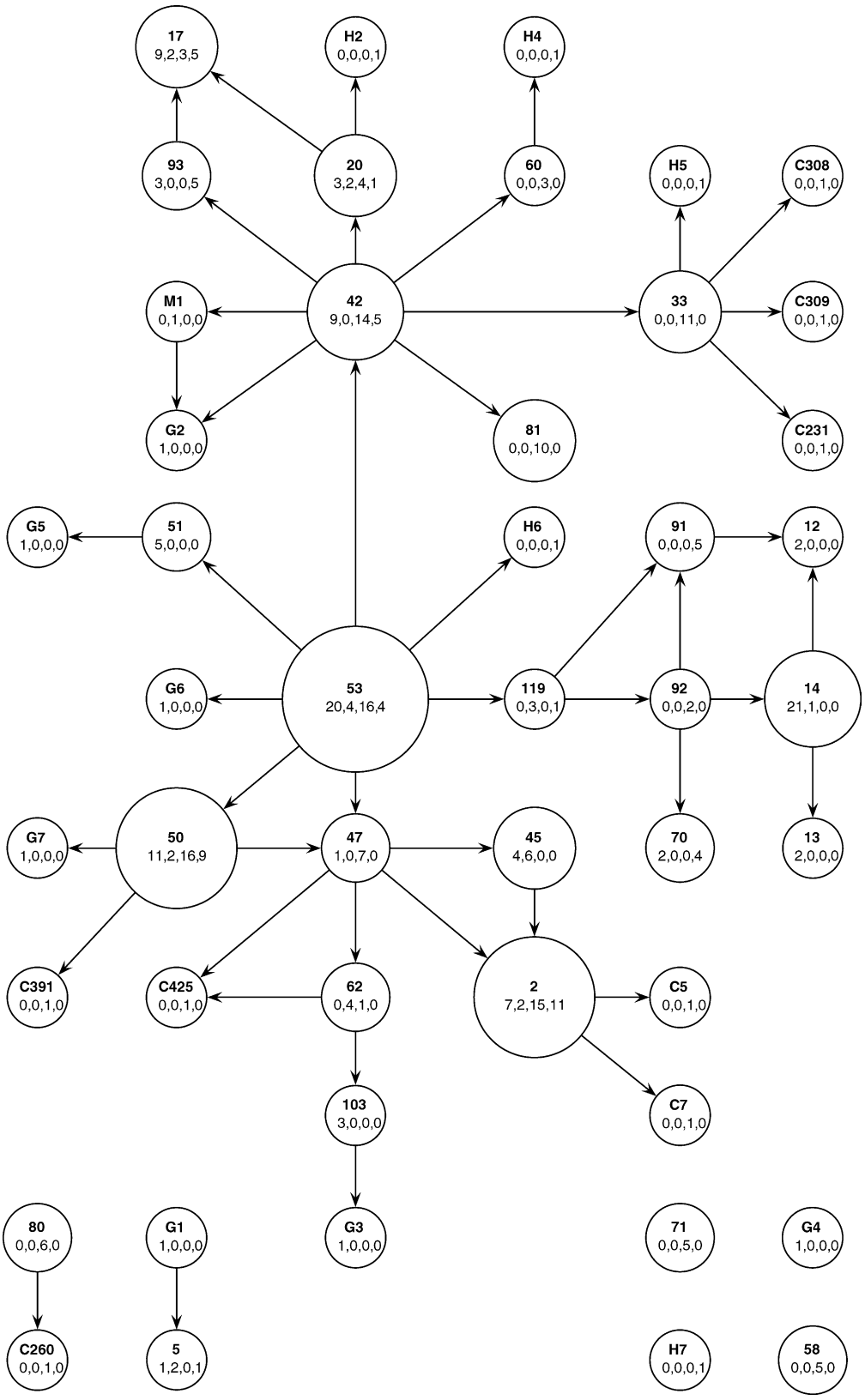

Fig. 3. Cluster-graph drawn from spoligotype data in Duchene et al. (2004). Vertex size indicates the size of the cluster; the list of fav numbers indicates the distribution of cases across the four regions from which isolates were taken (in order, Guadeloupe, Martinique, Haiti, Cuba). Arrows point to genotypes possibly derived by deletion in the marker locus.

we use the value $\mu = 0.04$ per year, a choice explained in Appendix C.

4. Graph-theoretic interpretation of indices

We now turn to the interpretation of some of the clustering indices described in this paper from a graph-theoretic point of view. If we construct an arbitrary transmission-forest subject to the constraints imposed by the cluster-graph of a given data set, then:

$$\mathrm{RTI}_{n-1} = \frac{\text{edges within clusters}}{\text{vertices in forest}}$$

and

$$\mathrm{RTI}^*_{n-1} = \frac{\text{edges within clusters}}{\text{maximum possible edges in forest}}.$$

Here, we see again that RTI_{n-1} must be strictly less than one, since the maximum number of edges in a forest is the number of edges in a tree spanning all vertices, which is one less than the number of vertices $(n-1)$.

Since the edges represent transmissions, RTI_{n-1} and RTI^*_{n-1} are natural measures of the severity of an outbreak. One advantage of viewing clustering in terms of transmission trees and forests is that the focus is shifted from cases to transmissions, which are the phenomena of interest.

Another example is Simpson's index. This is the proportion of all possible edges between vertices that are within clusters. Similarly, using Cayley's result that the number of rooted spanning trees on k vertices is k^{k-1}, we can interpret C_H (based on Shannon's index) as follows. Let σ be the number of forests consisting of trees spanning distinct clusters in the data set, and π be the number of ways of choosing one vertex from each cluster. Then

$$C_H = \frac{\ln(\sigma\pi)}{\max(\ln(\sigma\pi))},$$

where the maximum in the denominator is over all possible configurations of the data set into clusters (which occurs when all isolates are in the same cluster).

The TMI can also be interpreted in terms of transmission-forests:

$$\mathrm{TMI} = \frac{\text{mutation rate} \times \text{edges in forest}}{\text{edges involving mutation}}.$$

5. Discussion

We have examined some clustering indices for measuring the extent of transmission of clonal pathogens such as tuberculosis. A common problem of all of these indices, including those derived from diversity measures, is their failure to account for mutation. One possible solution is to prescribe guidelines for the interpretation of data based on the number of mutation steps apparently involved among the genotypes (Tenover et al., 1995). More quantitative and biologically explicit approaches may be possible in the future. One possibility is to derive indices that account for mutation, such as the TMI. As noted above (Section 2.3), the TMI defined in this paper is sensitive to some parameters, and requires further study.

Motivated by the difficulties arising from current methods, we have investigated an alternative way to summarise genotypic data using graph-theoretic concepts. We first introduced cluster-graphs, which are useful in visualising data. They provide information about the abundance of particular genotypes in the sample as well as partial information about possible evolutionary relationships. Rather than attempting to find the correct set of evolutionary relationships, we believe it may be more useful to show (non-tree) graphs that include alternative possibilities. Second, we discussed transmission trees, which *are* strictly trees. The idea behind these is to translate information about isolates and transmissions into graph-theoretic constructs, which can then be used as a way of understanding indices measuring the severity of outbreaks. Cluster-graphs and transmission trees may themselves suggest further ways to characterise disease transmission using a given data set.

Challenges remain in molecular epidemiology to the development of quantitative methods that (1) accurately reflect the severity of an outbreak, (2) account for mutation, and (3) are not susceptible to the effects of sampling. In addition to the theoretical development of such a method, an empirical problem is the determination of mutation rates of genetic markers. Although some progress has been made in recent years (Rosenberg et al., 2003), this programme is at an early stage. While one of the points of the present paper is that increased efforts should be made to estimate the mutation rates of markers, even a "rough idea" of the rates may be useful to know, as indicated in Table 1.

The effect of sampling is a universal problem in the analysis of molecular epidemiological data (as for all biological data). A number of studies examine the bias in estimates of clustering produced by sampling (Glynn et al., 1999; Murray, 2002). These papers find that the incomplete nature of all samples systematically leads to the underestimation of clustering. In general, the relationship between sample size n and the number of observed genotypes g is complex and depends on the details of both sampling and mutation as well as population history. A further related issue is the manner in which data are collected: random samples are very different from samples resulting from contact tracing, which are different again from samples collected at a given point such as a hospital or clinic. Future work should address the issue of sampling, possibly by adapting population genetic approaches (Ewens, 1972).

Based on this investigation, our main recommendations for the molecular epidemiology of tuberculosis and other diseases are as follows.

1. Caution should be taken when interpreting the transmissibility of outbreaks based on indices such as $\mathrm{RTI}_{n-1}, \mathrm{RTI}_n, S$ and C_H. It is perilous to compare outbreaks using different markers or different combinations of markers without an adequate framework that accounts for mutation. Similarly, data sets of different size should only be compared with caution.
2. In practice, it will be useful to identify relationships among genotypes through single mutation steps. The results can then be presented in graphs (trees) such as in Feil et al. (2004), Zhu et al. (2001), and Duchene et al. (2004) or cluster-graphs such as Fig. 1. Graphs (phylogenies or the other graphs discussed here) do not indicate by themselves the severity of an outbreak, but they do assist in organising the data to make sense visually.

Acknowledgements

We thank N. Rosenberg and R. Lan for their helpful comments. This work was supported in part by a Faculty Research Grant from the University of New South Wales.

Appendix A. Transmission trees

To use transmission trees as a model of epidemics, we describe the necessary biological assumptions.

1. Each individual in the sample was infected once.
2. Each individual in a given connected component of the cluster graph is connected via transmission to another in the same connected component.
3. Each genotype can arise only once. This is the same assumption giving rise to the Infinite Alleles Model (IAM) in population genetics.
4. Although the mutation rate is high enough to be modelled simultaneously with transmission, we will assume it is low enough that at most, only a single mutation event can occur during a given passage through a host. This assumption was imposed earlier by considering only single-step mutations in the cluster-graph, and can readily be relaxed if the marker is known to mutate rapidly.

The first two assumptions imply that all cases whose isolates are sufficiently closely related are potentially connected by transmission. Within a given cluster of size n_i, there are exactly $n_i - 1$ transmissions that took place among the cases in the cluster (a tree of n vertices has $n - 1$ edges). These two assumptions are inspired by reasoning in (Small et al., 1994).

Let each edge between clusters in a cluster-graph represent a single potential transmission event. These transmissions are therefore associated with mutation events. The infinite alleles assumption (number 3) together with assumption 1 imply that there can be at most one edge between clusters and that the data can only be explained by *spanning trees* of the connected components.

Appendix B. An index based on the maximum likelihood estimation of passage time

Define the parameter τ as the time between transmission events. In other words, τ is the time the pathogen has to undergo changes while in the transmitting or receiving host. Another way to think of τ is as the length of each edge in a transmission tree. Let μ be the mutation rate per unit time of the genetic marker being studied, where time is measured in the same units as τ.

Let M be the random variable describing the number of changes in a given time interval. Under the common assumption that genotypes change according to a molecular clock (Felsenstein, 1981; Tavare et al., 1997; Rosenberg et al., 2003),

$$P(M = m) = \frac{(\mu\tau)^m \, \mathrm{e}^{-\mu\tau}}{m!}.$$

We make the further assumption that mutation is rare enough that in a given transmission there can be at most a single change in a genotype. In other words, we are only concerned with $P(M = 0)$ and $P(M = 1)$.

Let E be the set of edges in the transmission tree. For an edge $j \in E$ let m_j be the (minimal) number of mutation steps connecting the two genotypes associated with the edge (the restriction on the number of mutations in a transmission means that we are assuming $m_j = 0$ or 1). The likelihood of the compound parameter $\mu\tau$ is the probability of observing the genotypes in the data set under the model of mutation:

$$\begin{aligned}\mathrm{Lik}(\mu\tau) &= P(\text{Genotypes}|\mu\tau) \\ &= \prod_{j\in E} P(M = m_j) \\ &= \prod_{j\in E,\, m_j=0} \mathrm{e}^{-\mu\tau} \prod_{j\in E,\, m_j=1} \mu\tau \, \mathrm{e}^{-\mu\tau}.\end{aligned}$$

Define ν_1 to be the number of edges involving a 1-step change, and ν_{2+} to be the number of edges involving multi-step changes. Setting $\nu_{2+} = 0$ (we have assumed $m_j = 0$ or 1) the logarithm of the likelihood is then $\ln(\mathrm{Lik}(\mu\tau)) = -\mu\tau(n - g + \nu_1) + \nu_1 \ln(\mu\tau)$, where g is the number of genotypes and n is the total number of cases in the data set. Therefore, the maximum likelihood estimator (MLE) for $\mu\tau$ is $\widehat{\mu\tau} = \nu_1/(n - g + \nu_1)$. Note that the number ν_1 is uniquely defined for a given cluster-graph as each spanning *cluster-tree* will have the same number of 1-step mutations. The number ν_1 is also equal to the number of genotypes g in the data set minus the number of connected components.

We can use the maximum likelihood estimate $\widehat{\mu\tau}$ to construct a measure of the speed of an epidemic that is inversely related to the passage time τ:

$$\mathrm{TMI} = \frac{\tilde{\mu}}{\widehat{\mu\tau}} = \tilde{\mu}\frac{(n-g+\nu_1)}{\nu_1}, \tag{B.1}$$

where $\tilde{\mu}$ is an independent estimate of the mutation rate.

Appendix C. Mutation rate of spoligotypes

A working estimate of the mutation rate of spoligotypes can be made by comparing the levels of spoligotype diversity with genotypes from IS*6110*, for which the mutation rate is known (Rosenberg et al., 2003).

Ewens (1972) showed that under a standard population genetic model (the infinite alleles model) the maximum likelihood estimator of an important population parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate, is given by the solution of

$$g = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i}$$

where g is the observed number of alleles in a sample of genotypes.

In the globally sampled data of Kremer et al. (1999) 84 different genotypes (alleles) out of 90 *M. tuberculosis* complex isolates were observed using IS*6110*. With the same isolates, 61 genotypes were observed using spoligotyping. The population conditions are identical for these two sets of numbers since they come from the *same* 90 isolates. Hence, the *relative* mutation rate of spoligotypes and IS*6110* can be estimated by taking the ratios of the estimates of θ. Using Ewens' sampling theory, these figures imply that spoligotypes mutate at around 13.5% the rate of IS*6110*. Using an earlier estimate of IS*6110* mutation rate (Rosenberg et al., 2003) of 0.287 per genome for a strain with a typical copy number of 10, this corresponds to a mutation rate for spoligotypes of around 0.039 events per year.

References

Alland, D., Kalkut, G., Moss, A., McAdam, R., Hahn, J., Bosworth, W., Drucker, E., Bloom, B., 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. New Engl. J. Med. 330 (24), 1710–1716.

Bauer, J., Andersen, A., Kremer, K., Miorner, H., 1999. Usefulness of spoligotyping to discriminate IS*6110* low-copy-number *Mycobacterium tuberculosis* complex strains cultured in Denmark. J. Clin. Microbiol. 37 (8), 2602–2606.

Coenye, T., Spilker, T., Martin, A., LiPuma, J.J., 2002. Comparative assessment of genotyping methods for epidemiologic study of *Burkholderia cepacia* genomovar III. J. Clin. Microbiol. 40 (9), 3300–3307.

Dale, J., Al-Ghusein, H., Al-Hashmi, S., Butcher, P., Dickens, A., Drobniewski, F., Forbes, K., Gillespie, S., Lamprecht, D., McHugh, T., Pitman, R., Rastogi, N., Smith, A., Sola, C., Yesilkaya, H., 2003. Evolutionary relationships among strains of *Mycobacterium tuberculosis* with few copies of IS*6110*. J. Bacteriol. 185 (8), 2555–2562.

Duchene, V., Ferdinand, S., Filliol, I., Guegan, J.F., Rastogi, N., Sola, C., 2004. Phylogenetic reconstruction of *Mycobacterium tuberculosis* within four settings of the Caribbean region: tree comparative analyse and first appraisal on their phylogeography. Infect. Genet. Evol. 4 (1), 5–14.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor Popul Biol 3 (1), 87–112.

Excoffier, L., Smouse, P., 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136 (1), 343–359.

Excoffier, L., Smouse, P., Quattro, J., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131 (2), 479–491.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P., Spratt, B.G., 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 186 (5), 1518–1530.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Foxman, B., Riley, L., 2001. Molecular epidemiology: focus on infection. Am. J. Epidemiol. 153 (12), 1135–1141.

Glynn, J.R., Vynnycky, E., Fine, P.E., 1999. Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. Am. J. Epidemiol. 149 (4), 366–371.

Hunter, P., Gaston, M., 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 26 (11), 2465–2466.

Kremer, K., van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W., Martin, C., Palittapongarnpim, P., Plikaytis, B.B., Riley, L.W., Yakrus, M.A., Musser, J.M., van Embden, J.D., 1999. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. 37 (8), 2607–2618.

Murray, M., 2002. Sampling bias in the molecular epidemiology of tuberculosis. Emerg. Infect. Dis. 8 (4), 363–369.

Niemann, S., Richter, E., Rusch-Gerdes, S., 1999. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J. Clin. Microbiol. 37, 409–412.

Rhee, J.T., Tanaka, M.M., Behr, M.A., Agasino, C.B., Paz, E.A., Hopewell, P.C., Small, P.M., 2000. Use of multiple markers in population-based molecular epidemiologic studies of tuberculosis. Int. J. Tuberc. Lung Dis. 4 (12), 1111–1119.

Rosenberg, N.A., Tsolaki, A.G., Tanaka, M.M., 2003. Estimating change rates of genetic markers using serial samples: applications to the transposon IS*6110* in *Mycobacterium tuberculosis*. Theor. Popul. Biol. 63 (4), 347–363.

Seidler, A., Nienhaus, A., Diel, R., 2004. The transmission of tuberculosis in the light of new molecular biological approaches. Occup. Environ. Med. 61 (2), 96–102.

Selander, R., Levin, B., 1980. Genetic diversity and structure in *Escherichia coli* populations. Science 210 (4469), 545–547.

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., Schoolnik, G.K., 1994. The epidemiology of tuberculosis in San Francisco: A population-based study using conventional and molecular methods. New Engl. J. Med. 330, 1703–1709.

Soini, H., Pan, X., Teeter, L., Musser, J.M., Graviss, E.A., 2001. Transmission dynamics and molecular characterization of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110*. J. Clin. Microbiol. 39 (1), 217–221.

Tanaka, M.M., Rosenberg, N.A., 2001. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Stat. Med. 20 (16), 2409–2420.

Tavare, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145 (2), 505–518.

Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Persing, D.H., Swaminathan, B., 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. 33 (9), 2233–2239.

Toungoussova, O.S., Sandven, P., Mariandyshev, A.O., Nizovtseva, N.I., Bjune, G., Caugant, D.A., 2002. Spread of drug-resistant *Mycobacterium tuberculosis* strains of the Beijing genotype in the Archangel Oblast, Russia. J. Clin. Microbiol. 40 (6), 1930–1937.

van der Spuy, G.D., Warren, R.M., Richardson, M., Beyers, N., Behr, M.A., van Helden, P.D., 2003. Use of genetic distance as a measure of ongoing transmission of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 41 (12), 5640–5644.

Yeh, R.W., Ponce De Leon, A., Agasino, C.B., Hahn, J.A., Daley, C.L., Hopewell, P.C., Small, P.M., 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J. Infect. Dis. 177 (4), 1107–1111.

Zhu, P., van der Ende, A., Falush, D., Brieske, N., Morelli, G., Linz, B., Popovic, T., Schuurman, I., Adegbola, R., Zurth, K., Gagneux, S., Platonov, A., Riou, J., Caugant, D., Nicolas, P., Achtman, M., 2001. Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. Proc. Natl. Acad. Sci. USA 98 (9), 5234–5239.