ELSEVIER

Phylogenetic Analysis as a Tool in Molecular Epidemiology of Infectious Diseases

BARRY G. HALL AND MIRIAM BARLOW

Phylogenetics is a powerful tool for microbial epidemiology, but it is a tool that is often misused and misinterpreted by the field. Microbial epidemiologists are cautioned that in order to draw any inferences about the order of descent from a common ancestor it is necessary to correctly root a phylogenetic tree. Epidemiological samples of microbial populations typically include both ancestors and their descendants. In order to illustrate the relationships of those isolates, the phylogenetic method used must be able to detect zero-length branches. Unweighted Pair-Group Method (UPGMA) is the phylogenetic method that is most widely used in microbial epidemiology. Because UPGMA cannot detect zero length branches, and because it places the root of the tree based on a usually-false assumption, UPGMA is the worst possible choice among the several phylogenetic methods available. Because microbial epidemiology deals with relationships among strains within a species, rather than with relationships among species, recombination within those species can render phylogenetic trees meaningless and positively misleading. When there is evidence of significant recombination within the species of interest phylogenetic trees should not be used at all. Instead, alternative tools such as eBURST should be used to understand relationships among isolates. *Ann Epidemiol 2006;16:157–169.* © 2006 Elsevier Inc. All rights reserved.

KEY WORDS: Phylogenetic Trees, UPGMA, Microbial Epidemiology, Recombination, eBURST.

INTRODUCTION

When conducting an epidemiologic study, the goal of that study is generally to determine the cause or the source of some health related phenomenon that affects a population and the distribution of that phenomenon throughout a population (1). While the source of many chronic, behavioral, or noninfectious diseases can be determined by studying the attributes, behaviors, and environment of the population of interest, it is often much more difficult to track the source of an infectious disease within a population by using these methods. The difficulties in tracking the source of an infectious agent occur because the pool of individuals infected by the disease experiences turnover of infected individuals (2), clinical laboratories have limited resources for identifying and reporting cases (2, 3), numerous infectious diseases cause similar symptoms (4), and many infected people do not seek treatment for their infections (5).

For over 30 years, molecular epidemiology has served as a very important tool for studying the spread of infectious diseases (4). Restriction fragment length polymorphisms (RFLP), randomly amplified polymorphic DNA (RAPD), and more recently, multiple locus sequence typing (MLST) have been used to determine the relatedness of bacterial strains (4). While these methods are useful in identifying whether there is a single source or if there are multiple sources of an infectious agent, by themselves, these methods are limited in their ability to identify the source of an infectious strain because they do not tell us anything about the direction in which evolution has occurred. For example, if a noninfectious strain is closely related to an infectious strain, none of these methods tells us whether the infectious strain was derived from the noninfectious strain or vise versa.

Phylogenetic methods can be used to analyze nucleotide sequence data, such as those that are available in MLST analyses in such a way that the order of descent of related strains can be determined. When coupled with appropriate phylogenetic analysis, molecular epidemiology has the potential to elucidate mechanisms that lead to microbial outbreaks and epidemics. Despite the utility of phylogenetics and the inexpensive, readily available software and manuals available for phylogenetic analyses, phylogenetic methods are often inappropriately applied. Even when appropriately applied, they are often poorly explained and are therefore poorly understood. Because phylogenetic analysis is inexpensive, especially when sequence data are already available, and because phylogenetic analysis shows much more clearly how infectious agents are spreading and evolving than sequence data alone, it is important for molecular epidemiologists to understand, to correctly apply, and to correctly interpret phylogenies and phylogenetic methods. This review, while not comprehensive is intended

From the Biology Department, University of Rochester, Rochester, NY (B.G.H.); the Bellingham Research Institute, Bellingham, WA (B.G.H.); and the Department of Epidemiology, Rollins School off Public Health, Emory University, Atlanta, GA (M.B.).

Address correspondence to: Barry G. Hall, Bellingham Research Institute, 218 Chuckanut Point Rd., Bellingham, WA 98229. E-mail: drbh@mail.Rochester.edu.

Received March 24, 2005; accepted April 20, 2005.

© 2006 Elsevier Inc. All rights reserved.
360 Park Avenue South, New York, NY 10010

1047-2797/06/$–see front matter
doi:10.1016/j.annepidem.2005.04.010

AEP Vol. 16, No. 3
March 2006: 157–169

Selected Abbreviations and Acronyms

RFLP = restriction fragment length polymorphisms
RAPD = randomly amplified polymorphic DNA
MLST = multiple locus sequence typing
RFLP = restriction fragment length polymorphisms
PFGE = pulsed field gel electrophoresis
ST = sequence type

to give molecular epidemiologists an overview of the methods, uses, and interpretations of phylogenetic trees derived from MLST data. Although we will discuss MLST data, the same discussion will apply equally well to viral sequence data and to data obtained by analysis of whole-chromosome restriction fragments that have been separated by pulsed field gel electrophoresis (PFGE).

There are several reasons that, since its introduction in 1998 (6), MLST has become the method of choice for microbial epidemiology. Because the data are DNA sequences the method is reproducible from laboratory to laboratory. Because an MLST scheme for any particular organism involves a defined set of primers for amplifying the DNA sequences, data from different laboratories are directly comparable and can be pooled into a single ever expanding data set that is stored in a single database. The data can be accessed over the Internet (http://www.mlst.net/databases/default.asp), and there are currently MLST databases for 20 microbial pathogens available, with another five schemes in development. The automation of DNA sequencing and the declining cost of automated DNA sequencing have made MLST a practical tool for epidemiologists. Finally, MLST typically has more resolving power than other methods, and is therefore preferable to earlier methods.

Typically, MLST data involves partial sequences, 400–600 base pairs, of several (typically 6–8) housekeeping genes that are dispersed around the bacterial chromosome. The sequence variation between two alleles of a locus is usually in the range of 0.1%–5%. Housekeeping genes are chosen because, being essential to life, they are under moderate to intense purifying selection. As a result, most of the sequence variation is the result of synonymous nucleotide substitutions that are close to selectively neutral. Because neutral variation accumulates approximately linearly with time (the molecular clock), genetic distance between alleles tends to be proportional to the time between divergence of those alleles. Because epidemiologists are usually interested in identifying and tracking pathogenic clones, some of which will have evolved only recently from nonpathogenic ancestors, it is important to have sufficient sequence variation to distinguish clones from close relatives. That level of variation is achieved by sequencing 6–8 loci. The loci are chosen to be well separated and scattered roughly evenly about the chromosome in order to assess the contribution of recombination to the variation as discussed later in the section on the importance of recombination in evaluating phylogenetic relationships.

The data are analyzed by assigning to each unique sequence of a locus an allele number. The *allelic profile* for an individual is the series of integers that represent the allele numbers at each of the loci, and each unique allelic profile defines a *sequence type* (ST). Individuals that have the same ST are identical at all of the loci examined and are presumed to be clones unless other reliable data (serotype, pathogenicity, metabolic properties) distinguish them.

APPLICATION OF PHYLOGENETIC METHODS TO EPIDEMIOLOGY

For the sake of this review, let us assume that we have a sample that includes many isolates of a pathogen that is responsible for a disease outbreak. The pathogen might be a microorganism or it might be a virus. In the former case, our raw data will probably be the complete or partial sequences of several genes (MLST); in the latter case, it will be the complete or partial sequences of the viral genome. In either case, we want to know how those isolates are related to each other, how they are related to their common ancestor, and how they have changed as they diverged from that common ancestor. A comparison of the nucleotide sequences in the form of a phylogenetic tree will greatly aid a correct understanding of the relationships among those organisms. Integration of that phylogenetic information with other information can help understand how the disease spread. For instance, are geographically nearby isolates more closely related to each other than are geographically distant isolates?

There are three primary MLST sites, each consisting of several databases and related software: the MLST home page at http://www.mlst.net/, the PubMLST home page at http://pubmlst.org/ and the MLST databases at the MPI für Infektionsbiologie home page at http://web.mpiib-berlin.mpg.de/mlst/. The MLST home page includes links to the other two pages. Software for MLST analysis, including allele assignment, allelic profile determination, assignment to STs, and construction of distance matrices based on pairwise nucleotide differences in the STs, can be downloaded from http://pubmlst.org/software/. The START program, available from that site, can carry out all stages of handling MLST data from sequence entry, through clustering, to drawing a UPGMA (unweighted pair-group method) tree. The relationships among isolates are typically determined by UPGMA from the matrix of ST distances using programs such as MEGA (http://www.megasoftware.net/) (7), PAUP* (http://paup.csit.fsu.edu/) (8) or PHYLIP (http://evolution.genetics.washington.edu/phylip.html) (9),

Download English Version:

https://daneshyari.com/en/article/3445810

Download Persian Version:

https://daneshyari.com/article/3445810

Daneshyari.com