APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 1996, p. 4299–4301 Vol. 62, No. 11
0099-2240/96/$04.00+0
Copyright © 1996, American Society for Microbiology

Problems in Measuring Bacterial Diversity and a Possible Solution

MILIND G. WATVE* AND RAJEEV M. GANGAL
Life Research Foundation, Navi Peth, Pune 411 030, India

Received 27 November 1995/Accepted 1 September 1996

The indices of species diversity used by plant and animal ecologists are not appropriate for bacterial diversity because of the inherent difficulty of defining a bacterial species. Arbitrary cutoff points to define a species or biotype lead to severe statistical problems. We suggest in this paper that a mean dissimilarity-based index without any attempt to define a species provides a statistically sound measurement of bacterial diversity.

Species diversity has been a key concept in ecological literature. A large number of diversity indices have been suggested and used by ecologists and mathematicians (e.g., references 12 and 14). Relatively few microbial ecologists have used quantitative measures of microbial diversity. The situation is, however, changing rapidly, with the number of papers making use of diversity indices in microbial ecology increasing considerably (e.g., references 3, 10, 11, 13, 16–18). Most microbial ecologists have borrowed diversity indices from plant and animal ecologists, the most popular ones being the Shannon index and Simpson's index (12) of diversity. There is a basic difficulty, however, in applying these indices to bacterial communities. These indices need a clear definition of species and an unambiguous identification of each individual, both of which are difficult in bacteriology. These difficulties have been acknowledged (16, 18), but the use of these indices continues, probably for the lack of better alternatives.

Studies of bacterial diversity differ from each other in the way of obtaining isolates, mode of characterization, clustering methods used for grouping or identification, the level of similarities or distances used to define a species or biotype, and the diversity measures used. Diversity measured by one method of characterization cannot be readily compared with that measured by other modes of characterization. We have not addressed such problems here. In this paper, we restrict our attention only to the statistical problems associated with bacterial diversity measurements and suggest alternatives. We first lay down a set of criteria for an ideal diversity index, test the commonly used diversity indices against these criteria, and then suggest an alternative approach to diversity measurement which satisfies these conditions and which is more suitable for bacterial communities.

Framework for selection. An ideal diversity index for bacterial communities should satisfy the following conditions, although it would be difficult for a single index to equally satisfy all.

(i) The index should reflect three important dimensions of diversity, viz. (i) the species richness or the number of different biotypes, (ii) their relative abundances, and (iii) the differences or the taxonomic distances between the biotypes.

(ii) The index should not be based on an arbitrary choice of any parameter, such as the similarity level at which a group of isolates is recognized as a species or a biotype. If such a parameter is essential, it should be statistically justified, and the index should not be sensitive to small changes in this parameter.

(iii) The index should not be disproportionately sensitive to possible errors and variability of test results.

(iv) Since samples of microbial communities are necessarily incomplete, very small in comparison with the size of the ecosystem, and destructive (9), the index should not be unreasonably sensitive to the sample size.

Indices of species richness and rarefaction curves are inadequate and inappropriate for bacterial communities. Simpson's index, the Shannon index, and evenness indices give enough weighting to the relative abundances along with species richness but fail to take into account the taxonomic distances between species, individuals, or any other appropriate unit. An information-based index would treat a community of four different biotypes of coliforms identical to another community consisting of one species of coliforms, one of actinomycetes, one of myxobacteria, and one of archaebacteria, whereas we feel that the latter should be treated as more diverse. The index therefore should consider the taxonomic distances. This concept is not new and has been addressed by mathematicians (14).

Both species richness and relative abundance indices are based on correct identification to the species level, which is seldom possible in bacteriology. Therefore, instead of attempting species identification, most researchers subject the characterization data to cluster analysis and identify groups of isolates that lie close to each other. In order to identify groups, it is often necessary to choose a cutoff level of similarity above which isolates fall in one group. This cutoff point is arbitrary and most often appears without a justification. For example, with physiological characterization, cutoff points of 80% (18), 75% (2), and 60% (6) similarities and 0.005 U of Ch12 taxonomic distance (3) have been used. With restriction fragment length polymorphism patterns for characterization, Leach et al. (11) used 85% similarity as a cutoff point, whereas van Berkum et al. (19) used 70% similarity as a cutoff point. With fatty acid methyl ester analysis, Boehm et al. (4) used a similarity index of ≥ 0.5 to define a species, whereas Ka et al. (8) used a similarity index of ≥ 0.6 to identify a species.

The choice of a cutoff point is crucial in species-based diversity measurements. The number of species as well as species-based indices would decrease by lowering the cutoff similarities. This decrease is nonlinear and often unpredictable, as shown by an analysis of four different sets of data (Fig. 1). If distinct flat regions are obtained in this curve, one can choose a cutoff point at the center of the flat portion. Because of the local plateau, small changes in the cutoff point or errors and

* Corresponding author. Mailing address: Life Research Foundation, 1000/6-C, Navi Peth, Pune 411 030, India. Phone: 0212 438905.

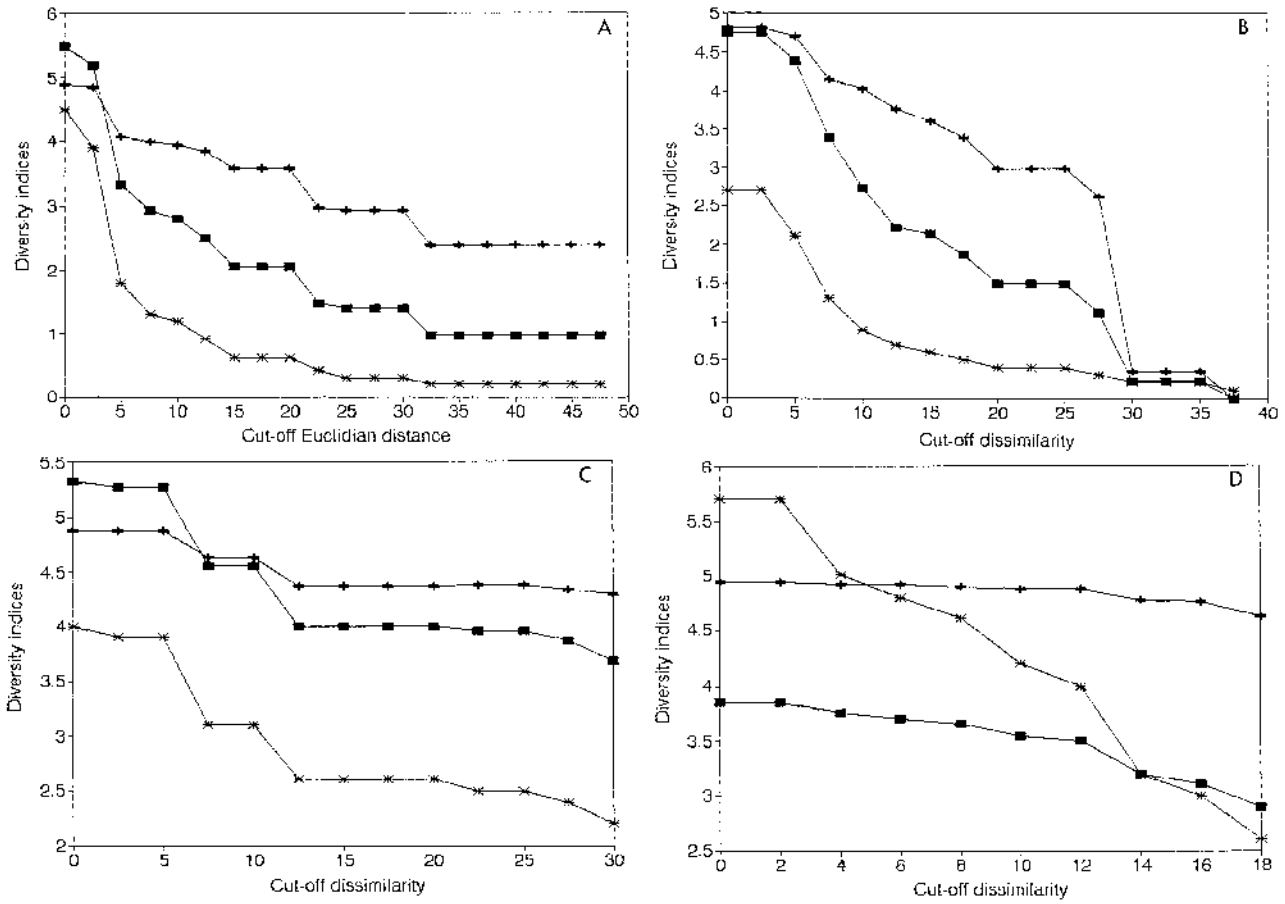

FIG. 1. Effect of the choice of the cutoff dissimilarity value on the diversity measures with four sets of data. If the cutoff point lies in a steep portion of the curve, the indices become highly sensitive to errors. The four sets of data differ in the locations and extent of steep and flat regions. Simpson's index (+) is less sensitive to a change in cutoff point than are the Shannon index (■) and species richness (∗). (A) Data from Leach et al. (11). (B) Data from Ka et al. (8). (C) Data on 40 isolates of *Acinetobacter* sp. from Dhakephalkar (5). (D) Data on earthworm gut flora from Marathe and Patil (12a).

variabilities of test results which affect the similarity values will not affect the diversity index. All sets of data, however, may not give distinct plateaus. Furthermore, the positions of plateaus in each set of data could be different, and therefore this approach is unlikely to give a generalizable solution. It may be possible to reach an agreement for an arbitrary but universally accepted cutoff point. This may apparently lead to a standard definition of a species, but for the communities in which this cutoff lies in a steep region of the curve, test errors and variabilities can lead to disproportionately larger errors in the diversity index, thus violating the third condition. The cutoff points of Ka et al. (8) and Leach et al. (11), for example, do not lie on plateaus, although they are not in the steepest portion of the curve. The use of cutoff similarities to define species thus suffers from several statistical problems.

Diversity without species: a novel approach. Since most of the problems in measuring bacterial species diversity arise from the difficulties of defining species, diversity indices which do not need a species or any other taxonomic level as a unit should be preferred. A simple alternative is to use the mean taxonomic distance (*D*mean) between all pairs of isolates as a diversity index (14):

$$D\text{mean} = \frac{\Sigma\, d(i,j)}{S(S\ +\ 1)/2}$$

where $d(i, j)$ is the taxonomic distance between the ith and jth isolate and S is the total number of isolates. Distance or dissimilarity can be defined as (1 − Ssm) or (1 −SJ) for binary characterization data, where Ssm is the simple matching coefficient and SJ is the Jaccard coefficient of similarity (15). Euclidean distance, $1 - r$ (where r is the correlation coefficient between test scores of two isolates), or 1 − nucleic acid homologies can also be used. The mean dissimilarity determined by any of the above methods reflects all of the three dimensions of diversity. It generally increases with increasing number of distinct biotypes and decreases if one biotype dominates the community, thus reflecting both richness and relative abundances. It reflects the taxonomic distances directly and does not involve any arbitrary parameter. Dissimilarity measures are thus promising candidates as bacterial diversity indices.

Dissimilarity coefficients are, however, not free of problems. A large mean dissimilarity can be obtained by a small number of distantly related biotypes or a large number of moderately related ones. For example, a bipolar community having only two biotypes which are equally abundant and completely dissimilar will have a mean (1 − Ssm) of 0.5. A hypothetical community generated by randomizing binary test results with equal probabilities of positive and negative results also gives a mean (1 − Ssm) of 0.5. This is counterintuitive. A bipolar community with only two diametrically opposite types should

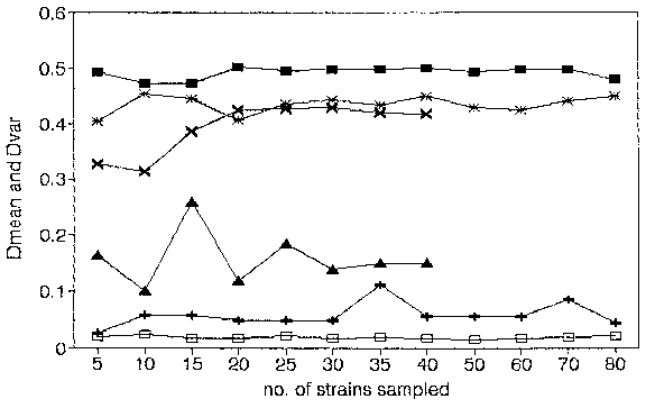

FIG. 2. Sample size invariance of *D*mean and datum variance (*D*var) according to Jaccard dissimilarities. Both of the values remain stable after a sample size of 20. Symbols for sets of data are as follows: ■, set 1 mean; +, set 1 variance; ∗, set 2 mean; □, set 2 variance; ×, set 3 mean; ▲, set 3 variance. Data for set 1 were obtained from P. Dhakephalkar (5). Data for sets 2 and 3 were computer-generated fictitious data.

be much less diverse than the randomized results community in which most of the isolates would have a unique set of characteristics.

The problem can be resolved easily. Although the bipolar and the randomized communities show identical mean Ssm, the bipolar community has much higher variance around the mean. The diversity index therefore should include the mean dissimilarity along with the variance. The mean provides the diversity measure per se, and the variance gives additional useful information about the community. A large mean accompanied by small variance indicates a large number of moderately dissimilar biotypes. A large mean accompanied by large variance can be interpreted as a community with a few dominant biotypes that are taxonomically distinct. Dominance of a single biotype will result in a small mean and small variance.

The dissimilarity-based indices with binary data are sensitive to the proportion of positive results, since the proportion decides the probabilities of positive and negative matches. The simple dissimilarity coefficient (1 − Ssm) is maximum at equal proportions of positive and negative results and decreases as the ratios become skewed in either direction. The Jaccard dissimilarities decrease monotonically with the proportion of positive results, since this index ignores the negative matches. A normalization of the index against the effect of the proportion of positive results is therefore needed. This can be achieved by dividing the empirical mean dissimilarity by the expected dissimilarity of a randomized community at the given proportion of positive results.

When using (1 − Ssm), a normalized index would be

$$D = \frac{D\text{mean}}{[a^2 + (1 - a)^2]}$$

where a is the fraction of positive tests. Similarly, with (1 − SJ) the normalized index becomes

$$D = \frac{D\text{mean}}{a^2/[1 - (1 - a)^2]}$$

Both of the mean dissimilarity indices show sample size invariance (Fig. 2) for $n > 20$ when tested against hypothetical as well as factual data. The dissimilarity-based measures of diversity thus satisfy all of the conditions laid down at the beginning to a greater or lesser extent. Expression of all of the information of a community in a single number is essentially a reduction in information, and therefore no single index reflects all aspects of diversity. However, the approach suggested here is particularly suitable for bacterial communities and has several merits which the indices borrowed from eukaryotic communities lack.

REFERENCES

1. **Austin, B., and F. Priest.** 1986. Modern bacterial taxonomy. Van Nostrand Reinhold, London.
2. **Batzli, J. M., W. R. Graves, and P. van Berkum.** 1992. Diversity among rhizobia effective with *Robinia pseudoacacia* L. Appl. Environ. Microbiol. **58:**2137–2143.
3. **Bianchi, M. A. G., and A. J. M. Bianchi.** 1982. Statistical sampling of bacterial strains and its use in bacterial diversity measurement. Microb. Ecol. **8:**61–69.
4. **Boehm, M. J., L. V. Madden, and H. A. J. Hoitink.** 1993. Effect of organic matter decomposition level on bacterial species diversity and composition in relationship to pythium damping-off severity. Appl. Environ. Microbiol. **59:** 4171–4179.
5. **Dhakephalkar, P.** Personal communication.
6. **Frederickson, J. K., D. L. Balkwill, J. M. Zachara, S.-M. W. Li, F. J. Brockman, and M. A. Simmons.** 1990. Physiological diversity and distributions of heterotrophic bacteria in deep Cretaceous sediments of the Atlantic Coastal Plain. Appl. Environ. Microbiol. **57:**402–411.
7. **Haldeman, D. L., and P. S. Amy.** 1993. Diversity within a colony morphotype: implications for ecological research. Appl. Environ. Microbiol. **59:**933–935.
8. **Ka, J. O., W. E. Holben, and J. M. Tiedje.** 1993. Genetic and phenotypic diversity of 2,4-dichlorophenoxyacetic acid (2,4-D)-degrading bacteria isolated from 2,4-D-treated field soils. Appl. Environ. Microbiol. **60:**1106–1115.
9. **Kinkel, L. L., E. V. Nordheim, and J. H. Andrews.** 1992. Microbial community analysis in incompletely or destructively sampled systems. Microb. Ecol. **24:**227–242.
10. **Kuhn, I., G. Allestam, T. A. Stenström, and R. Möllby.** 1991. Biochemical fingerprinting of water coliform bacteria, a new method for measuring phenotypic diversity and for comparing different bacterial populations. Appl. Environ. Microbiol. **57:**3171–3177.
11. **Leach, J. E., M. L. Rhoads, C. M. Vera Cruz, F. F. White, T. M. Mew, and H. Leung.** 1992. Assessment of genetic diversity and population structure of *Xanthomonas oryzae* pv. oryzae with a repetitive DNA element. Appl. Environ. Microbiol. **58:**2188–2195.
12. **Ludwig, J. A., and J. F. Reynolds.** 1988. Statistical ecology: a primer on methods and computing, p. 85. John Wiley & Sons, New York.

12a. **Marathe, B., and S. Patil.** Personal communication.

13. **Moyer, C. L., F. C. Dobbs, and D. M. Karl.** 1994. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl. Environ. Microbiol. **60:**871–879.
14. **Rao, C. R.** 1980. Diversity and dissimilarity coefficients: a unified approach. Theor. Pop. Biol. **21:**24–43.
15. **Sneath, P. H. A.** 1972. Computer taxonomy. Methods Microbiol. **1972:**57–58.
16. **Staley, J. T.** 1980. Diversity of aquatic heterotrophic bacteria, p. 321–322. *In* D. Schlesslinger (ed.), Microbiology—1980. American Society for Microbiology, Washington, D.C.
17. **Torsvik, V., J. Goksøyr, and F. L. Daae.** 1989. High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. **56:**782–787.
18. **Torsvik, V., K. Salte, R. Sørheim, and J. Goksøyr.** 1989. Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. Appl. Environ. Microbiol. **56:**776–781.
19. **van Berkum, P., S. I. Kotob, H. A. Basit, S. Salem, E. M. Gewaily, and J. S. Angle.** 1993. Genotypic diversity among strains of *Bradyrhizobium japonicum* belonging to serogroup 110. Appl. Environ. Microbiol. **59:**3130–3133.