Commentary

How many species of prokaryotes are there?

Bess B. Ward*

Geosciences Department, Princeton University, Princeton, NJ 08544

The microorganisms classified in the two prokaryotic domains of the tree of life, Bacteria and Archaea, possess immense metabolic diversity, and their activities are critical in processes ranging from sewage treatment to regulating the composition of the atmosphere. Especially in light of the rate of modern climate change, it is essential to understand how microbial communities affect ecosystem functioning and how human activities, such as agriculture, waste management, and climate modification, affect microbial communities. Thus discovering and understanding the diversity of microbial communities (the number of species and their relative abundances) is a high priority in ecology. In this issue of PNAS, Curtis *et al.* (1) address what is therefore one of the most fundamental questions in microbial ecology—how many species of prokaryotes are there in nature?

Only in the last 10–15 years has it even been possible to pose the question and hope realistically for an answer in the case of prokaryotes. Less than 30 years ago, the answer to the even more fundamental question ''How many individuals are there?'' was revised in such a way as to change the entire focus of environmental microbiology. So to appreciate the significance of the question and answer provided by Curtis *et al.* (1), it is useful to review that recent history briefly.

Before the mid-1970s, microbial ecologists assessed the population size of bacteria in soils, sediments, and natural waters by culturing the microbes and counting the number of colonies that grew on nutrient agar plates. For seawater, the cultivable prokaryotic population size was a few hundred cells per milliliter (2, 3). That was an almost inconsequential number relative to the thousands or tens of thousands of planktonic algal cells that could be seen (literally, with a microscope) in the same milliliter of water. The primary importance then ascribed to environmental bacteria was their potential as pathogens, and research focused on survival of pathogens in natural environments. But when researchers saw millions of cells per milliliter, or per gram of soil or sediment, with electron microscopy (4) and epifluorescence microscopy [using DNA-binding fluorochromes such as acridine orange and later 4′,6-diamidino-2-phenylindole (DAPI)] (5, 6), they realized that bacteria must have much broader and potentially more important roles in natural systems. Those additional cells could not all be pathogens, but microbial ecologists were stymied in their efforts to identify them by the very fact that they did not grow on plates. So the answer to the question ''How many are there?'' led to intense interest in ''Who are they?'' and ''What are they doing?''

The answer to ''What are they doing?'' was provided by sensitive stable isotope and radiotracer methods: prokaryotes play essential roles in the primary production and consumption of organic matter and the cycling of nutrient elements in the modern environment and were perhaps even more important in the evolution of the atmosphere and hydrosphere before the appearance of eukaryotes.

The answer to ''Who are they?'' became infinitely more approachable with the advent of molecular biological methods in environmental microbiology. Woese (7) first recognized the utility of the ribosomal RNA molecule, with its universal distribution, and its high conservation coupled with moderate variability, for constructing global phylogenies of all living things. He also introduced the now widely accepted three-domain tree of life (8), in which prokaryotes constitute two branches (Bacteria and Archaea) and Eukarya the third. On the basis of a few hard-won sequences, the first ''universal'' primers for use in the polymerase chain reaction (PCR) were designed and used to pluck the 16S ribosomal RNA genes of uncultivated organisms right out of the soil and water of their natural environments. Once those genes were sequenced, their evolutionary relationships to cultivated and other uncultivated organisms could be determined. ''Who are they?'' became synonymous with ''Where do they fall in the 16S rRNA-based tree of life?'' Diversity surveys based on cultivation had identified 10 or 12 major divisions within the Bacteria and two or three in the Archaea (9). There are now more than 40 major bacterial divisions recognized (10) and 12 or more in the Archaea (11). The tree is still growing and not just at the twig level—just last month (May 2002), a new phylum of Archaea was discovered at a hydrothermal vent, its rRNA sequence so different from known groups that it could not be detected with the ''universal'' probes (12). Perhaps the most astounding finding, and one that will keep microbial ecologists busy for many years, is that most of the sequences retrieved from the environment without cultivation are not represented by any cultivated organisms. Entirely new microbial worlds have always been out there, beneath our feet, in our water and air. Many remarkable and fascinating new organisms have been discovered (on the basis of their 16S rRNA sequence) in extreme environments such as hydrothermal vents and hot springs, desert sand, and Antarctic oceans, but the smallest drop of temperate seawater or a grain of agricultural soil will also yield myriad 16S rRNA sequences that are new to science.

As these data on the immense diversity of life began to accumulate, the question that Curtis *et al.* (1) attempt to answer became unavoidable. ''How many different kinds can there possibly be?'' If with every milliliter of water or gram of sediment we discover new prokaryotic sequences, is there any limit to the diversity in nature? If we cannot enumerate all of the different sequences in one sample, how can we compare that sample to another to ask whether the community compositions of the two samples are different?

The calculation of community diversity by any conventional diversity index requires two fundamental pieces of information: the number of species and the number of individuals in each species. These data are often presented in a species abundance curve in which the number of species is plotted vs. the number of individuals per species. Although microbial ecologists have made much progress toward identifying large numbers of species, the quantitative information on how many of each is present is still an almost impossible goal. Curtis *et al.* (1) show that by assuming a log-normal distribution (i.e., most species have an intermediate number of individuals and few species have very small or very large populations), the area under the curve can be used to estimate the number of species from the total number of individuals. It is then possible to relate diversity of prokaryotic communities to the ratio of two potentially

See companion article on page 10494.

*E-mail: bbw@princeton.edu.

www.pnas.org/cgi/doi/10.1073/pnas.162359199

measurable variables: the total number of individuals (N_T) and the abundance of the most abundant species (N_{max}). Using the epifluorescence microscopy methods mentioned above, it is relatively easy to obtain an estimate of N_T (although there is current controversy over the possibility that a large fraction of the total may not be active or even alive). It then remains to determine N_{max}, and Curtis *et al.* (1) use information from two sources: direct counts using fluorescence *in situ* hybridization (FISH, using probes derived from the 16S rRNA sequence) and abundance of distinct sequences in clone libraries of 16S rRNA sequences obtained from the environment. FISH assays have been done for only a few species and it is not possible to know *a priori* which ones are likely to be most abundant, so FISH databases are quite limited. Clone libraries are biased by the initial PCR and are rarely sequenced to exhaustion. In addition, clone libraries based on 16S rRNA genes cannot distinguish between multiple copies of the *rrnb* operons from one cell or many from different cells of the same species. Since the number of operons can vary from 1 to 14 per genome, clone library representation cannot represent species number accurately. Thus clone libraries are not reliable indicators of N_{max} and it is difficult to anchor the species abundance distribution at all.

At the present time therefore, even the basic requirement for these two fundamental pieces of information cannot be met for most environments and most kinds of microbes. Curtis *et al.* (1) therefore based some of their quantitative comparisons on reports of diversity among communities of ammonia-oxidizing bacteria (AOB). The AOB are a useful group for this purpose because they are largely monophyletic on the basis of 16S rRNA and, at least compared with some other functionally defined groups, they are not very diverse. Monophyly means that, unlike many clusters in the tree of life, we can use the 16S rRNA sequence to identify the functional group and can, with a high degree of certainty, conclude that a 16S rRNA sequence in the AOB cluster does indeed come from an organism that oxidizes ammonia for a living. Curtis *et al.* (1) estimated the total species richness of AOB to be as low as 6 for the entire Arctic Ocean (13), and predicted a much larger number from AOB clone libraries from agricultural soils (14). Their obligate chemoautotrophic lifestyle apparently constrains both the diversity and abundance of AOB. By every measure, AOB constitute a miniscule fraction of the total bacteria in most environments, so despite their attractions as a model group, it may be difficult or inappropriate to extrapolate from them to the entire community.

Hughes *et al.* (15) approached the problem of "counting the uncountable" from the practical perspective of the researcher desiring to compare the species richness of multiple communities or sites. Then the question is "How many clones must be analyzed to (*i*) predict the total number present in each site and (*ii*) reduce the uncertainty of the estimate sufficiently to allow comparisons among estimates?" Hughes *et al.* (15) concluded that there was insufficient information to assume that microbial populations are log-normally distributed and therefore used nonparametric estimators, which depend on mark and recapture type methods to estimate the total number of species in the sample. That is, they use information on the proportion of species that have been observed before (more than once in a clone library, in this case) relative to those that are observed only once. The goal of their study was not explicitly to arrive at estimates of the total number of species (or operational taxonomic units, OTUs) in particular environments, but rather to demonstrate the utility of statistical approaches for doing so. Nonetheless, the examples they used are useful for comparison with the approach of Curtis *et al.* (1). For clone libraries from two grazed grassland soils (16), Hughes *et al.* (15) obtained estimates of 467 and 590 OTUs (which are not statistically different from each other). In analyzing the same clone data, Curtis *et al.* (1) obtained an estimate of 6,300, an order of magnitude greater.

Using a different species estimator (DNA–DNA association; see below), others have derived estimates of 350–1,500 OTUs in arable or metal-polluted soils (17), 6,000–10,000 (18) or up to 500,000 (19) in unperturbed soils. The importance of these numbers is that they are large. So large that determining the actual number of species (or OTUs) is not feasible empirically (by brute force exhaustive sequencing). So large that we can be confident we are still missing major parts of the extant diversity even in common settings such as agricultural soil. In addition, the range in the numbers is so unconstrained as to provide no support for the assumption of a log-normal—or any other—distribution of abundance.

The entire discussion about how many species there are presupposes that we know what a species is. Although the species is usually defined in evolutionary and genetic terms and in fact is subject to much debate, microbial ecologists are in addition deprived of the simple expedient of being able to tell species apart by looking at them. Morphology may not suffice to distinguish congeneric diatoms or worms, but it is even more woefully inadequate for prokaryotes. The same molecular methods that allowed microbial ecologists to ask "Who are they?" provided the means to distinguish one kind from another in the first place. The two most commonly used criteria today depend on (*i*) DNA–DNA association (i.e., the degree of hybridization possible between the total genomic DNA of two species) and (*ii*) the percent identity in the sequence of that ubiquitous molecule, 16S rRNA. Curtis *et al.* (1) simply accept that a meaningful distinction between kinds of organisms can be defined on the basis of 16S rRNA sequence, and most of their fellow researchers would agree in principle. Many might even agree on the conventional cutoffs of greater than 70% DNA–DNA reassociation or greater than 97% 16S rRNA identity (20) as species definitions.

16S rRNA sequence information allows the construction of phylogenetic trees that show the ancestry and relatedness of organisms on the basis of that molecule, but even organisms that are identical or cluster tightly by that criterion cannot be concluded to share all or, in some cases, essential physiological similarities. Thus definition of species on this basis is not adequate for assessing the functional diversity of prokaryotic communities. The discovery of ecologically important differences in temperature optima attributed to hot spring microbes with less than 1% 16S rRNA sequence divergence led Ward (21) to advocate a more "natural" or ecological species concept (ESC) of Simpson (22). The ESC includes information not just on genetics but on the role of the organism in its environment, some acknowledgment of its function. If one really wishes to understand the effect of diversity on ecosystem function, then a species definition that relates to species function is essential. Again, the AOB provide a useful example. In addition to their monophyletic distribution in terms of 16S rRNA, another advantage is that among that group, there is apparently only one way to oxidize ammonia. All of the AOB have a homologous gene (*amo*) encoding the enzyme ammonia monooxygenase, and the sequence of this gene can also be used to investigate the diversity of AOB. Much of the DNA sequence diversity in functional genes may not be selective (i.e., may not translate to differences at the protein level because of third codon wobble), and there is no agreed-upon cutoff for what constitutes a distinction at the DNA sequence level for functional genes. For the sake of this example, we will accept that 5% sequence difference at the DNA level constitutes a significant difference. Then for the studies summarized in Table 1, it can be seen that functional gene diversity usually exceeds that detected at the 16S rRNA level among AOB in several environments, and the trend is consistent even on the basis of small clone libraries derived from different sets of PCR primers.

Sequence diversity among functional genes generally exceeds that in the ribosomal genes on the basis of data from cultivated strains, and some highly conserved protein-encoding genes such as DNA gyrase

Table 1. Comparison of number of different kinds of AOB present in various environments

Environment	16S rRNA: No. of clones >3% divergent/total clones in analysis	*amoA*: No. of clones >5% divergent/total clones in analysis	Ref.
Sequencing batch biofilm reactor	5/21	5/12	27
Chalk grasslands, The Netherlands	4/15	11/NA*	28
Seawater, Monterey Bay, CA†	2/57	15/93	G. O'Mullan and B.B.W., unpublished data
Apr. 1998; 40 m	1/9	1/11	
Oct. 1998; 32 m	1/10	3/13	
Apr. 1999; 32 m	1/9	3/8	
Oct. 1999; 7 m	1/10	5/20	
Oct. 1999; 32 m	2/9	3/21	
Oct. 1999; 50 m	1/10	8/20	

Data are from clone libraries except where noted.

***amoA* diversity was assessed by analysis of terminal restriction fragment length polymorphism (TRFLP), which did not require cloning and sequencing, so the total number of *amoA* sequences retrieved cannot be estimated.

†First line: combined data from six clone libraries of each gene. Data for individual clone libraries identified by sampling date and depth are in the list below. Some groups occurred in more than one library (e.g., one rRNA group was present in all samples), so the number of different types in individual libraries does not sum to the number of different types overall.

(23, 24) and RNA polymerase (25) have been used for phylogeny. However, this direct comparison, which is so useful for AOB, cannot be made for most kinds of organisms in the environment because of the lack of correspondence between ribosomal phylogeny and functional genes. An extreme example of the problem is found in denitrifying microbes: the ability to denitrify occurs in all three domains of life but occurs sporadically among cultivated members of coherent 16S rRNA clusters (26). The functional genes involved in the denitrification pathway (e.g., nitrite reductase, nitrous oxide reductase) exhibit huge sequence diversity in the cultivated representatives and in a gram of marine sediment or forest soil. In this case 16S rRNA phylogeny is not informative about function, and counting the number of species in a 16S rRNA clone library would not shed much light on the functionality of the ecosystem.

A species is not defined by the sequence of one functional gene even if, like *amo*, it is the quintessential gene that defines the organism's metabolism. However, the organism's interaction with the environment—its regulation by environmental variables such as temperature, oxygen, and substrate concentrations—is defined at the level of functional genes and the enzymes they encode, not 16S rRNA. These are the genes that determine the role of the species, that comprise the essence of an ecological species concept. The number of functional genes for which sizeable databases are available is very small, but from these data it is obvious that the diversity of functional genes far exceeds that of the ribosome. If the diversity of functional genes reflects potential diversity in actual ecological function, then this diversity has implications for ecosystem function, resilience, and stability. In assessing the diversity of prokaryotic communities, some recognition of this additional layer of complexity must be included.

Whether considering rRNA, genome, or functional gene databases, the number of different kinds of prokaryotes is so large as to make counting them an endless task, and the problem may even be worse than Curtis *et al.* (1) estimate. There are essentially no databases yet available that are adequate to provide realistic estimates of the simple parameters required by the model, or even to constrain the shape of the species abundance curve. The contribution of Curtis *et al.* (1) is important in recognizing the size of the problem, even if it cannot yet answer the question of the magnitude of prokaryotic diversity. Technical advances occur rapidly in the field of molecular biology, and it may not be long before the data necessary to test the assumptions of the model are available. Microbial ecologists should follow the advice of Curtis *et al.* (1) and reject "the counsel of despair"; we should use these results to guide future experimental design, to bring forward the day when this fundamental question can be answered.

I thank Greg O'Mullan and Chris Francis for much critical discussion.

1. Curtis, T. P., Sloan, W. T. & Scannell, J. W. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 10494–10499.
2. Carlucci, A. F. & Pramer, D. (1957) *Proc. Soc. Exp. Biol. Med.* **96,** 392–394.
3. Jannasch, H. W. & Jones, G. E. (1959) *Limnol. Oceanogr.* **4,** 128–139.
4. Watson, S. W., Novitsky, J. T., Quinby, H. L. & Valois, F. W. (1977) *Appl. Environ. Microbiol.* **33,** 940–946.
5. Hobbie, J. E., Daley, R. J. & Jasper, S. (1977) *Appl. Environ. Microbiol.* **33,** 1225–1228.
6. Porter, K. G. & Feig, Y. S. (1980) *Limnol. Oceanogr.* **25,** 943–948.
7. Woese, C. R. & Fox, G. E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5088–5090.
8. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
9. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
10. Pace, N. R. (1997) *Science* **276,** 734–740.
11. DeLong, E. F. & Pace, N. R. (2001) *Syst. Biol.* **50,** 471–478.
12. Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C. & Stetter, K. O. (2002) *Nature (London)* **417,** 63–67.
13. Bano, N. & Hollibaugh, J. T. (2000) *Appl. Environ. Microbiol.* **66,** 1960–1969.
14. Bruns, M. A., Stephen, J. R., Kowalchuk, G. A., Prosser, J. I. & Paul, E. A. (1999) *Appl. Environ. Microbiol.* **65,** 2994–3000.
15. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. (2001) *Appl. Environ. Microbiol.* **67,** 4399–4406.
16. McCaig, A. E., Glover, L. & Prosser, J. I. (1999) *Appl. Environ. Microbiol.* **65,** 1721–1730.
17. Ovreas, L. (2000) *Ecol. Lett.* **3,** 236–251.
18. Torsvik, V., Daae, R. L., Sandaa, R. A. & Ovreas, L. (1998) *J. Biotechnol.* **64,** 53–62.
19. Dykhuizen, D. E. (1998) *Antonie Leeuwenhoek* **73,** 25–33.
20. Stackebrandt, E. & Goebel, B. M. (1994) *Int. J. Syst. Bacteriol.* **44,** 846–849.
21. Ward, D. M. (1998) *Curr. Opin. Microbiol.* **1,** 271–277.
22. Simpson, G. G. (1961) *Principles of Animal Taxonomy* (Columbia Univ. Press, New York).
23. Yamamoto, S. & Harayama, S. (1998) *Int. J. Syst. Bacteriol.* **48,** 813–819.
24. Venkateswaran, K., Moser, D. P., Dollhopf, M. E., Lies, D. P., Saffarini, D. A., MacGregor, B. J., Ringelberg, D. B., White, D. C., Nishijima, M., Sano, H., *et al.* (1999) *Int. J. Syst. Bacteriol.* **49,** 705–724.
25. Palenik, B. & Swift, H. (1996) *J. Phycol.* **32,** 638–646.
26. Zumft, W. G. (1997) *Microbiol. Mol. Biol. Rev.* **61,** 533–616.
27. Gieseke, A., Purkhold, U., Wagner, M., Amann, R. & Schramm, A. (2001) *Appl. Environ. Microbiol.* **67,** 1351–13562.
28. Kowalchuk, G. A., Stienstra, A. W., Heilig, G. H. J., Stephen, J. R. & Woldendorp, J. W. (2000) *Environ. Microbiol.* **2,** 99–110.