Estimating bacterial diversity from clone libraries with flat rank abundance distributions

Mary Lunn,1* William T. Sloan2 and Thomas P. Curtis3
1Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.
2Department of Civil Engineering, University of Glasgow, Oakfield Avenue, Glasgow G12 8LT, UK.
3School of Civil Engineering and Geosciences, University of Newcastle upon Tyne, Newcastle, NE1 7RU, UK.

Summary

There are a number of parametric and non-parametric methods for estimating diversity. However all such methods employ either the proportional abundance of the most abundant taxon in a sample or require that a specific taxon is sampled more than once. Consequently, the available methods for estimating diversity cannot be applied to samples consisting entirely of singletons, which might be characteristic of some hyperdiverse communities. Here we present a non-parametric method that estimates the probability that a given number of unique taxa would be sampled from a community with a particular diversity. We have applied this approach to a well known data set of 100 unique clones from a sample of Amazonian soil (Borneman and Triplett (1997) Appl Environ Microbiol 63: 2647–2653) and determine the probability that this observation would be made from an environment of a given diversity. On this basis we can state this observation would be very unlikely ($P = 0.006$) if the soil diversity was less than $10^3$, and quite unlikely ($P = 0.6$) if the diversity was less than $10^4$, and probable ($P = 0.95$) if the diversity was about $10^5$. There are essentially no contestable assumptions in our method. Thus we are able to offer almost unequivocal evidence that the bacterial diversity, of at least soils, is very large and a method that may be used to interpret samples consisting entirely of singletons from other hyperdiverse communities.

Introduction

The extent of microbial diversity is still uncertain. However, most (Cho and Tiedje, 2000; Torsvik et al., 2002) but not all (Finlay, 2002) commentators believe that the diversity we observe is vastly exceeded by the diversity we cannot observe. As a consequence microbial ecology requires tools to infer the latter from the former. In the first instance methods have been adapted for the study of microbial communities from methods developed for other problem domains. These methods have been authoritatively reviewed by Bohannan and colleagues (Hughes et al., 2001; Bohannan and Hughes, 2003) and may be: parametric, in which a particular species abundance curve is assumed, or non-parametric, in which no particular distribution is assumed.

For example, the best known non-parametric methods are those of Chao (1984;1987) who estimates the minimum possible diversity in a system from the observed numbers of singletons and doubletons in a sample of a particular size. These methods determine the lowest number of different individuals consistent with the frequency at which a single organism is counted more than once. The great advantage of Chao’s work is that it is non-parametric and thus makes no assumption about the underlying distribution, the great disadvantage is that it only estimates the lower limit of the maximum possible diversity and requires doubletons to be present.

This contrasts with the method of Curtis et al. (2002), which is parametric (it assumes a lognormal distribution) and may be used to estimate the maximum possible diversity in a given environment or functional group. This method exploits the ratio of the total number of individuals and abundance of the most abundant species. An important aspect of this particular approach is that the calculations were based on parameters that microbial ecologists can measure. This method’s principal weaknesses are the assumption of a lognormal species abundance curve and a particular minimum abundance (or the canonical lognormal) it thus might tend to overestimate the extent of microbial diversity. Intriguingly, this method seems to agree with the estimates of Torsvik et al. (2002), based on DNA:DNA reassociation. This might not be because the assumptions of Curtis and co-workers are correct, but because the method itself is insensitive to modest departures from these assumptions.

However, until the definitive and authoritative experimental definition of microbial diversity has been undertaken in a number of environments we will not be able to know if bacterial species abundance curves approximate to a particular shape. Moreover, theory (Hubbell, 2001)
suggests that taxon abundance patterns may be a function of the size of the global community, the size and isolation of the local community and the level of taxonomic resolution chosen. Species may have a lognormal distribution, whereas genera might have a log-series (W. T. Sloan, M. Lunn, I. M. Head and T. P. Curtiss, submitted). Thus at present, those seeking to estimate the diversity of a particular community are probably wise to triangulate the estimates of diversity by using parametric and non-parametric methods. These estimates will be essential if we are to design interventions to estimate diversity authoritatively.

The methods of Chao and Curtis and co-workers both require, respectively, that a doubleton may be observed or the abundance of the most abundant organism ($N_{\text{max}}$) may be known or inferred (ideally using fluorescence in situ hybridization). In practice, workers use clone libraries and count the doubletons or the proportional abundance of the most frequent clone.

However, in very diverse communities such as soils the same sequence simply may not be sampled more than once and thus no obvious abundant clone and no means of estimating the unobserved diversity. For example, Borneman and Triplett (1997) prepared a clone library of 16S rRNA genes recovered from an Amazonian soil by PCR, which yielded 100 unique sequences. This situation is likely to be repeated wherever less conserved sequences are examined in more diverse environments.

Here we present a non-parametric method that allows the minimum possible diversity to be inferred from such flat distributions.

**Basic rationale**

Here we present our rationale which we believe to be intuitively credible. A detailed proof is presented in the appendix.

*Find what the diversity is not*

Estimating total diversity from a clone library is essentially a sampling problem. We wish to learn something about the source community from the frequency with which members of the source community are observed in a sample. We can approach this problem in two ways by either trying to work out what the source diversity is (which may be hard) or trying to work out what the diversity is not (which might be easier). The likelihood that the number of species in the source community is not less than a certain number will increase as more and more unique species are observed. This will of course be affected by the underlying distribution of the source community, for if the underlying distribution is very uneven and yet only unique clones are observed, then the prediction will be that the number of species is very high indeed. In practice, we cannot be sure what the underlying distribution is. However, if we calculate all probabilities assuming a flat distribution, we can be sure that our probabilities and associated estimates are conservative (see Appendix).

On this basis and assuming that we are sampling without replacement (a valid assumption because there are so many individuals in the microbial world) from a uniform distribution then the probability, that all the individuals in a sample of size $L$ are singletons, is

$$P(\text{all singletons}) = \frac{J^{-1}(S - 1)!}{S^{-1}(S - L)!J - 1!}$$

where, $J$ is the numbers of individuals in the community and $S$ is the number of species in the community.

This equation calculates the probability that every individual in the sample is a singleton (Fig. 1). Thus by inserting the observed numbers of singletons one can then estimate how many species there must be if the probability that only singletons are observed is at least as large as for example 0.5 (i.e. 50% chance of observing all singletons). On this basis we are able to revisit the intriguing dataset of Bornemann and Triplett. Assuming that there are $10^9$ individuals in the wider community, we have found it is extremely unlikely (probability $= 0.006$) that all singletons are observed in a sample of size 100 (Fig. 2) if there are less than $10^3$ species in the community sampled, there is a probability of 0.6 if diversity is about $10^4$ and a probability of 0.95 if the diversity is about $10^5$. Looking at this sample another way, we can say that $(1686, \pm 2$) is a 95% confidence interval for species diversity.

These estimates are consistent with, but larger than the estimates of Torsvik et al. (1990; 2002) and the estimates of Curtis et al. (2002). However, whereas the estimates of Torsvik and Curtis and co-workers pertained to a sample,
Torsvik’s work (Torsvik 1990) is still implicitly contested assumptions in this estimate (given that all the sequences recorded are indeed different). We cannot of course know the value of \( J \), however, we note that, in the expression for the probability given above, the contribution from \( J \) is virtually 1 provided that \( J \) is very large (say \( 10^8 \)). Similarly, if the distribution is not as flat as we assumed, \( S \) will be larger. Thus we are able to offer further, almost unequivocal evidence that the bacterial diversity of soil is, at least in some cases, very large indeed. The apparently unequivocal nature of our reasoning is important because Torsvik’s work (Torsvik et al., 1990) is still implicitly contested (Finlay, 2002).

The proposition that microbial diversity is small has arisen from work with protozoa. Many protozoan morphotypes appear to be ubiquitous and therefore it has been suggested that protozoan diversity is low. The rationale offered for this observation (that large scale dispersion across geographical barriers limits speciation and extinction) appears to apply to all microbial life: including the bacteria (Finlay, 2002). Interestingly, the local and global diversity of the protozoa appears to have a log series distribution (Finlay and Clarke, 1999; Finlay, 2002). In a study on the ubiquity of the protozoan genus *Paraphysomonas* Finlay and Clarke examined 1607 individuals from a sediment sample recently derived from a community of \( 3.8 \times 10^{14} \) individuals and observed 31 species. This compared well with a stated global diversity of 41–50 species. We used our approach to simulate this experiment but use a discretised log-series distribution (which is observed in practice) rather than a flat distribution, and assumed a global diversity of 50 in a local community of \( 10^{14} \). We found that we would expect to find 49.9 (variance 0.009) species in a sample of 1607 individuals. Thus our approach appears to work with low diversity and high diversity communities. This raises the intriguing question of why such radically different diversities are observed. An understanding of the distribution of bacterial taxa at local and global scales will be of fundamental strategic importance to our future understanding of the microbial diversity, irrespective of the extent of the microbial world. Recently published simulations illustrate the effect of underlying distribution on the rate of discovery of new sequences (Narang and Dunbar, 2003). One consequence of this insight is that the apparent decrease in the rate of discovery of new sequences in marine environments (Hagstrom et al., 2002) might reflect the abundant distribution of a minority of organisms rather than an absolute lack of diversity in the seas.

We suspect that the significance of flat distributions has probably been overlooked in the past and we believe that some clone libraries with flat rank abundance distributions have gone unpublished because it was thought they were not amenable to analysis. Furthermore when only a small number of doubletons are observed in a large clone library screened with amplified rDNA restriction analysis (ARDRA) or restriction fragment length polymorphism (RFLP) (Zhou et al., 2002) it may be well worthwhile sequencing the few putatively similar sequences to find out if they really are the same since a truly flat clone library will be of some significance.

We are now coming to realise that the size of the microbial community could be of profound practical importance (W. T. Sloan, M. Lunn, I. M. Head and T. P. Curtiss, submitted), explaining, for example, why small anaerobic reactors do not appear to be stable (Fernandez et al., 1999). Our work may also help inform those who wish to tackle the global genome question in soils. This is clearly an extremely difficult endeavour. At the present the authoritative definition of the microbial diversity of any community is challenging and fraught with profound technical and sampling problems (Dunbar et al., 2002). Given the cost and difficulty of measuring microbial communities, it may often be cheaper and quicker to adapt the mathematical tools to the microbes than to match the measuring of the microbes to the mathematical tools. This is a modest first attempt. However, we are also aware that some of the most promising tools (Orlitsky et al., 2003) seem very hard to understand.

**Acknowledgements**

We thank Bland Finlay for providing the sample size data from his *Paraphysomonas* work. We would like to thank our referees for helpful comments and discussion.
References


Chao, A. (1984) Nonparametric-estimation of the number of species, an individual is equally likely to be drawn from any species, i.e. a flat distribution. If J is very large, and is also large in comparison with S, then we can approximate using sampling without replacement,

\[
\Pr(\text{all singletons}) = 1 - \left( \frac{J}{J-1} \right)^{J} \left( \frac{J}{J-2} \right)^{J} \ldots \left( \frac{J}{J-L+1} \right)^{J} \]

where \(J - \frac{J}{J-1}\) is the probability that the second individual sampled is not the same species as the first and so on. Written more tidily,

\[
\Pr(\text{all singletons}) = \frac{J^{J} - 1}{S^{J}(S-L)} \frac{(J-L)}{(J-1)(J-2)) \ldots (J-L+1)}
\]

which is the probability of all singletons if the sampling is done with replacement.

1. Appendix

1.1 Probability of no repetition in the individuals sampled

We calculate the probability that all singletons are observed in a sample of size L from the wider community J with individuals and S species on the assumption that an individual is equally likely to be drawn from any species, i.e. a flat distribution. If J is very large, and is also large in comparison with S, then we can approximate using sampling without replacement,

\[
\Pr(\text{all singletons}) = 1 - \left( \frac{J}{J-1} \right)^{J} \left( \frac{J}{J-2} \right)^{J} \ldots \left( \frac{J}{J-L+1} \right)^{J} \]

where \(J - \frac{J}{J-1}\) is the probability that the second individual sampled is not the same species as the first and so on. Written more tidily,

\[
\Pr(\text{all singletons}) = \frac{J^{J} - 1}{S^{J}(S-L)} \frac{(J-L)}{(J-1)(J-2)) \ldots (J-L+1)}
\]

which is the probability of all singletons if the sampling is done with replacement.

1.2 Assuming a flat distribution gives a conservative estimate of the total number of species

Suppose that we wish to calculate the maximum expected number of species present in a sample, given that the number of individuals in the meta-community is so large that sampling with replacement is assumed. Suppose that S is the number of species in the meta-community and that \(p_i\) is the probability that a single microbe selected at random is from species i. Note that \(p_i \geq 0\) and \(\sum_{i=1}^{S} p_i = 1\).

Then we will show that the expected number of species in the sample is maximised if all species are equally likely, that is if we have a flat distribution.

Suppose the sample size is L and that the number of species represented in the sample is \(S_L\). Then

\[
E(S_L) = S - \sum_{i=1}^{S} E(l(i))\]  where \(l(i) = 1\), if the ith species is present, and \(l(i) = 0\) if it is absent.
And so
\[
E = E(S_i) = S - \sum_{i=1}^{S} (1 - p_i)^i
\]
To maximise \(E\) differentiate subject to the constraint \(\sum p_i = 1\)
\[
\frac{\partial E}{\partial p_i} = -L(1 - p_i)^{i-1} = \lambda, \quad i = 1,2,\ldots,S
\]
using a Lagrange multiplier \(\lambda\) (Bunday, 1984) and giving \(p_i = p_2 = \cdots = p_s = \frac{1}{S}\) as required. The flat distribution gives the maximum value of \(E(S_i)\). The minimum expected number of species is 1.

We can also see that the flat distribution gives the largest probability that a sample \((L < S)\) consists entirely of singletons, from the following calculation. Consider the probability that \(k\) species are present in a sample of size \(L\)
\[
P(S_i = k) = \sum_{i_1 + \cdots + i_k = L} \prod_{j=1}^{k} p_{i_j}^{i_j}, \text{summed over}
\]

\[
\Gamma(i_1) + \cdots + \Gamma(i_k) = k
\]
where
\[
\Gamma(i_j) = \begin{cases} 1 & \text{if } i_j \geq 1 \\ 0 & \text{if } i_j = 0 \end{cases}
\]
Now suppose that, \(k = L\), giving \(j = 1\) or 0. Set the function \(f\) to be
\[
f = P(S_i = L) = \prod_{i_1,\ldots,i_k : i_j = 0} \prod_{i_1,\ldots,i_k : i_j \geq 1} p_{i_j}^{i_j}
\]
Then we need to maximise \(f\) subject to the constraint \(\sum p_i = 1\) as before. Solve \(\frac{\partial f}{\partial p_i} = \lambda\) again using a Lagrange multiplier. Take \(i = 1,2\) for example, we see
\[
\frac{\partial f}{\partial p_1} = \prod_{i_1,\ldots,i_k : i_j < \delta} p_{i_j}^{i_j} - \prod_{i_1,\ldots,i_k : i_j \geq \delta} p_{i_j}^{i_j} = \lambda
\]
\[
\frac{\partial f}{\partial p_2} = \prod_{i_1,\ldots,i_k : i_j \geq \delta} p_{i_j}^{i_j} - \prod_{i_1,\ldots,i_k : i_j < \delta} p_{i_j}^{i_j} = \lambda
\]
After some algebra we see, on subtracting,
\[
(p_1 - p_2) \sum_{0 < k < \delta \leq S} p_k = 0.
\]
We deduce \(p_1 = p_2\) (since \(p_1 = 0\) for some \(j\) can only reduce the number of species). Similarly for all other probabilities \(p_i\) and we again have \(p_1 = p_2 = \cdots = p_s = \frac{1}{S}\)
A flat distribution gives the highest probability of selecting only singletons \((L < S)\).

1.3 Expectation and variance of number present in the sample size \(L\)
Suppose as before that \(p_i\) is the probability that a single microbe selected at random is from species \(i\). Then \(E(S_i) = S - \sum_{i=1}^{S} (1 - p_i)^i\) as above. We can also compute the variance of \(S_i\) using
\[
\text{var}(S_i) = \text{var} \left( \sum_{i=1}^{S} (1 - p_i)^i \right) = E \left( \sum_{i=1}^{S} (1 - p_i)^i \right)^2 - \left( E \sum_{i=1}^{S} (1 - p_i)^i \right)^2.
\]
Since
\[
E((i(i)) = (1 - p_i - p_i)^i, \quad i \neq j
\]
we must have
\[
\text{var}(S_i) = \sum_{i=1}^{S} (1 - p_i)^i + \sum_{i>j} (1 - p_i - p_i)^i - \left[ \sum_{i=1}^{S} (1 - p_i)^i \right]^2
\]
(as in Harris, 1959). The calculations for the flat distribution follow easily. If we wish to use the log series so that the number of species \(S_i\) with precisely \(i\) individuals is
\[
\frac{\delta}{i} \left( 1 - \frac{\delta}{i} \right)^i
\]
then the above formula can be simplified a little by grouping the species of size \(i\), numbering \(S(i)\) in total.