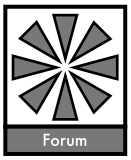

Oikos 117: 177–181, 2008
doi: 10.1111/j.2007.0030-1299.16311.x,

Subject Editor: Heikki Setälä, Accepted 2 November 2007

DNA reassociation kinetics and diversity indices: richness is not rich enough

Bart Haegeman, Dimitri Vanpeteghem, Jean-Jacques Godon and Jérôme Hamelin

B. Haegeman (bart.haegeman@inria.fr), D. Vanpeteghem, J.-J. Godon and J. Hamelin, INRA, UR050, Laboratory of Environmental Biotechnology, Avenue des Etangs, FR-11100 Narbonne, France. BH also at: INRA-INRIA MERE research team, UMR Systems Analysis and Biometrics, 2 place Pierre Viala, FR-34060 Montpellier, France.

DNA reassociation kinetics, also known as Cot curves, were recently used by Gans and co-workers to estimate the number of bacterial species present in soil samples. By reanalysing the mathematical model we show that rather than the number of species, Simpson and Shannon diversity indices are encoded in the experimental data. Our main tool to establish this result are the so-called Rényi diversities, closely related to Hill numbers, illustrating the power of these concepts in interpreting ecological data. We argue that the huge diversity encountered in microbial ecology can be quantified more informatively by diversity indices than by number of species.

Measuring microbial diversity is a challenging problem. First of all, the concept of microbial species is still under debate (Stackebrandt et al. 2002, Gevers et al. 2005). The definition used for eukaryotes (sexual compatibility) is indeed useless for bacterial and archaeal organisms. Next, a choice should be made from the wealth of available diversity notions (Magurran 2004). Richness, which takes all species equally into account, seems to be less appropriate due to the huge number of rare species in microbial communities. Shannon or Simpson diversity and in particular unifying concepts like Hill numbers (Hill 1973) look more promising. They give gradually stronger weighting to dominant species than to rare ones in quantifying diversity.

Although these diversity measures have been introduced long ago, microbial diversity estimation still focuses mainly on richness (Curtis et al. 2002, Venter et al. 2004, Gans et al. 2005, Hong et al. 2006, Loisel et al. 2006, Schloss and Handelsman 2006). Rare species, whose abundance is difficult to assess experimentally, can then be dealt with by assuming a species abundance distribution (SAD). However, the number of possible SADs is large (Magurran 2004), and it is unknown which are the realistic ones in a microbial context. Moreover, richness estimates often depend on this SAD assumption, and differ sometimes by orders of magnitude (Venter et al. 2004, Gans et al. 2005, Hong et al. 2006, Loisel et al. 2006).

This problem seems to be invariably present for different measurement techniques, and might deteriorate significantly the estimation precision. Moreover, microbial diversity estimation is also hindered by a number of intrinsic experimental biases. The vast majority of DNA-based techniques observe microbial communities through the 16S ribosomal RNA gene. This requires polymerase chain reaction (PCR) amplification, introducing a bias which is difficult to quantify (Forney et al. 2004). The PCR fragments are then analysed either by fingerprints or by molecular inventories. The former technique is rapid but imprecise (Loisel et al. 2006); the latter uses rarefaction to obtain diversity estimates, which poses a number of statistical problems (Lande et al. 2003, Hong et al. 2006, Schloss and Handelsman 2006). More recently, metagenomic approaches were used to assess microbial community complexity (Venter et al. 2004). This extremely heavy technique bypasses PCR amplification, but introduces an equally unknown bias in the cloning step.

Rather different is the method based on reassociation kinetics of single stranded DNA, because it avoids amplification biases. Initially developed to determine the amount of distinct DNA in a solution (Britten and Kohne 1968), it was later adapted to define phylogenetic relatedness between bacterial organisms, and has by now become the official technique to delineate bacterial species (Stackebrandt et al. 2002). The first application of reassociation kinetics in microbial ecology estimated bacterial richness of forest soil (Torsvik et al. 1990a, 1990b). The data analysis was recently extended to include SAD assumptions (Gans et al. 2005), but the error was shown to be much larger than the estimated richness (Bunge et al. 2006, Volkov et al. 2006).

However, we demonstrate here that accurate diversity information is encoded in reassociation profiles. We first introduce a technical tool, the Rényi diversity (Rényi 1961), and briefly recall some of its properties relevant for ecology

(Hill 1973). Next, we present a simple derivation of the reassociation kinetics model of Gans et al. (2005). A simulation study shows the tight connection between diversity indices and reassociation times. The fitting error to estimate diversity from experimental data has a clear structure, revealing that rather than the number of species, the diversity indices like Simpson's and Shannon's can be accurately estimated. Finally, we speculate what this result tells about the diversity concept in microbial ecology.

Rényi diversities

From a theoretical point of view, the analysis presented in this paper is based on a diversity notion introduced by Rényi (1961) in the context of information theory. Denoting by n_s the relative abundances, the Rényi diversities are defined by:

$$R_\alpha = \frac{1}{1-\alpha} \ln \sum_{s=1}^{S} n_s^\alpha$$

with $\alpha \geq 0$, where for $\alpha = 1$ the limit $\alpha \to 1$ is understood. Note that Rényi diversities are directly related to Hill numbers (Hill 1973) by $H_\alpha = \exp R_\alpha$.

Common diversity measures (Magurran 2004) like the richness S, the Shannon diversity index H,

$$H = -\sum_{s=1}^{S} n_s \ln n_s$$

the Simpson concentration index C,

$$C = \sum_{s=1}^{S} n_s^2$$

are contained in the Rényi diversities. Indeed,

$$R_0 = \ln S, \quad R_1 = H, \quad R_2 = -\ln C$$

Note that all these quantities depend on relative abundances n_s rather than absolute abundances N_s. One can prove that R_α is a decreasing function of α, and thus $\ln S \geq H \geq -\ln C \geq 0$.

This study will exploit randomly assembled communities. To generate them, we use a small program written in Matlab (The Mathworks Inc., Natick), that is available upon request. First, we choose the number of species S, by drawing $\log_{10} S$ uniformly in the interval [0,5]. Next, we select the absolute abundance N_s for every species s, independently and from the same species abundance distribution (SAD). As our analysis will only require relative abundances n_s, we finally divide the absolute abundances N_s by the total number of individuals $\Sigma_s N_s$.

We focus on the following SADs:

- The lognormal distribution, given by the probability density function

$$\rho_1(N) = \frac{1}{\sqrt{2\pi}\sigma N} \exp\left(-\frac{(\ln N - \mu)^2}{2\sigma^2}\right), \quad N \in]0, \infty[.$$

  Going from absolute to relative abundances eliminates the parameter μ. This can be proven by computing the probability density function for the set of relative abundances n_s. The parameter σ is drawn uniformly in [0,4].
- The power-law distribution, given by the probability density function

$$\rho_2(N) = \frac{z-1}{N^z}, \quad N \in [1, \infty[.$$

  The parameter z is drawn uniformly in [1.2, 3].
- The truncated power-law distribution with exponent $z = 1$, given by the probability density function,

$$\rho_3(N) = \frac{1}{aN}, \quad N \in [1, e^a].$$

  Since the function 1/N is not integrable over $[1, \infty[$, it is necessary to truncate the distribution for large N (at e^a in this case). The parameter a is drawn uniformly in [0,20].

We believe that this set of one-parameter SADs with the mentioned parameter ranges covers a broad range of realistic community structures. Finally, we mention the equal-abundance community, $n_s = 1/S$ for all s, that has been used in data interpretation of Cot curves (Torsvik et al. 1990a, 1990b).

A model for reassociation kinetics

To make our exposition self-contained, we present an elementary derivation of the model of reassociation kinetics introduced by Gans et al. (2005). Denote the concentration (mass/volume) of dissociated single stranded DNA molecules of species s at time t by $C_s(t)$. Using mass action kinetics, the reassociation is described by the differential equation

$$\frac{dC_s}{dt} = -kC_s^2$$

where we assume that the reassociation of species s is not modified by the presence of other species. Supposing all molecules are dissociated at the initial time $t = 0$, the solution of the differential equation is given by

$$C_s(t) = \frac{C_s(0)}{1 + kC_s(0)t}$$

To better fit experimental data from *Escherichia coli* DNA, one introduces an empirical parameter, the retardation factor γ,

$$C_s(t) = \frac{C_s(0)}{(1 + kC_s(0)t)^\gamma}$$

Note that both parameters k and γ are assumed to be the same for all species s. Our analysis will not depend on the numerical value of the rate constant k. For the retardation factor γ we take $\gamma = 0.45$ (Gans et al. 2005). The total concentration of dissociated molecules C(t) is given by

$$C(t) = \sum_{s=1}^{S} C_s(t) = \sum_{s=1}^{S} \frac{C_s(0)}{(1 + kC_s(0)t)^{\gamma}} = \sum_{s=1}^{S} \frac{n_s C(0)}{(1 + kn_s C(0)t)^{\gamma}} \quad (1)$$

with n_s the relative abundances. We arrived at the formula that Gans et al. (2005) used to describe reassociation curves.

The simplifying assumptions in the derivation of the DNA reassociation kinetic model are numerous. For example, species that contain repeated sequences (Britten and Kohne 1968, Godde and Bickerton 2006), or sequences shared by different species (Choi and Kim 2007) are not taken into account. Also, the parameters k and γ might differ from species to species (Volkov et al. 2006), which would complicate the model considerably. Finally, it has been pointed out that a realistic renaturation model should also include the dynamics after complementary strands have made contact (Murugan 2003, Bunge et al. 2006).

Linking reassociation and diversity

Our present aim is not to improve the DNA reassociation kinetic model, but to use Eq. 1 to clarify the diversity estimation problem. In Cot curve analysis of microbial communities, diversity has been interpreted exclusively as richness. However, the seminal paper of Torsvik et al. (1990a) suggested already a possible correlation with the Shannon diversity index. Rényi diversities and Eq. 1 allow us to systematically establish this link.

Our strategy is as follows. We randomly generate a large number of communities, for which we compute, on one hand, the Rényi diversities, and on the other, the Cot curve. Our goal is to correlate both. To do so, we introduce the reassociation times, defined as the time needed to renaturate a given fraction, say β, of the initial mixture of single stranded DNA molecules. We then establish a link between Rényi diversities and reassociation times. More precisely, we show that with any fraction β, we can associate an index α, such that the Rényi diversity R_α can be predicted (with small error) from the reassociation time for that fraction β. This indicates that the Cot curve model encodes a range of Rényi diversities.

Our argument starts by introducing the rescaled time $\tau = kC(0)t$ and the fraction $c(\tau)$ of dissociated molecules at time τ. We then rewrite Eq. 1 as

$$c(\tau) = \sum_{s=1}^{S} \frac{n_s}{(1 + n_s\tau)^{\gamma}}$$

The reaction rate k has been absorbed in the dimensionless time τ. The resulting function is monotonically decreasing from 1 to 0, so we can define times τ_β such that $c(\tau_\beta) = \beta$ for any $\beta \in]0, 1[$.

We fix a pair (α, β) and investigate the relationship between the Rényi diversity R_α and the reassociation time τ_β. For each of the 3 SADs (lognormal, power-law and truncated power-law with $z = 1$) we generate twice 1000 communities, and compute the Rényi diversity R_α and the reassociation time τ_β. The relation R_α vs $\ln \tau_\beta$ is fitted with a linear function on the first set of 1000 communities. Then, this function is used to predict R_α from τ_β for the second set of 1000 communities. The root mean squared error made in this prediction is denoted by $E(\alpha, \beta)$. We repeat this procedure for every pair (α, β).

Figure 1 shows the contour plot of the error function $E(\alpha, \beta)$. A curve of minimal error can be traced in the (α, β) plane (shown in thick line). On this curve the error is everywhere smaller than 0.1. It tends to $\beta = 1$ for the Simpson diversity $\alpha = 2$, but does not get to the species richness $\alpha = 0$, even for small β. For pairs (α, β) off this curve, the fitting error $E(\alpha, \beta)$ increases rapidly.

Horizontal sections at $\beta = 0.01$, $\beta = 0.5$ and $\beta = 0.99$ are shown in Fig. 2. The error $E(\alpha, \beta = 0.5)$ reaches a minimum for the Shannon diversity $\alpha = 1$. This corresponds to the time needed so that half of the DNA has reassociated, a quantity often used in Cot analysis (Britten and Kohne 1968, Torsvik et al. 1990a, 1990b). Similarly, the error for $\beta \approx 1$ has a minimum for the Simpson diversity $\alpha = 2$. The section at $\beta = 0.01$ has a minimum at $\alpha \approx 0.5$, and the error increases steeply for smaller α. Nevertheless, it is for small values β that the richness S has to be found. The logarithm ln S cannot be estimated with an error smaller than 1, indicating the poor correlation between species richness and reassociation times.

Figure 3 shows the correlation between R_α and $\ln \tau_\beta$ for three pairs (α, β), corresponding to the best estimates for logarithmic richness ln S, Shannon index H and Simpson index $-\ln C$. The errors are $E(\alpha = 0, \beta = 0.01) = 1.1$, $E(\alpha = 1, \beta = 0.5) = 0.07$ and $E(\alpha = 2, \beta = 0.99) = 0.02$, for diversities R_α that vary in the range 0 to 10. Thus, the simulations suggest that estimation of the Simpson diversity is more precise than the Shannon diversity. Moreover, the Simpson diversity can be obtained for shorter reassociation times, and thus smaller experimental errors.

An analytical expression for the link between R_α and $\ln \tau_\beta$ can be obtained by considering equal-abundance

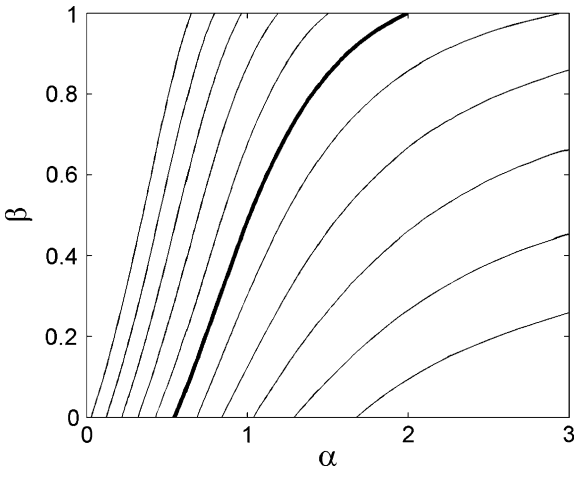

Fig. 1. Fitting error $E(\alpha, \beta)$ for reassociation kinetics. Twice 1000 communities were generated for each of the three SADs (lognormal, power-law and truncated power-law with $z = 1$). Rényi diversities R_α and reassociation times τ_β were computed. The first set of communities was used to construct a linear fit R_α vs $\ln \tau_\beta$; the second set to compute the corresponding fitting error. This error $E(\alpha, \beta)$ has a valley-like structure, the bottom of which is shown in thick line. The thin lines are, for increasing distances from the thick line, the 0.2, 0.4, 0.6, 0.8 and 1 level curves.

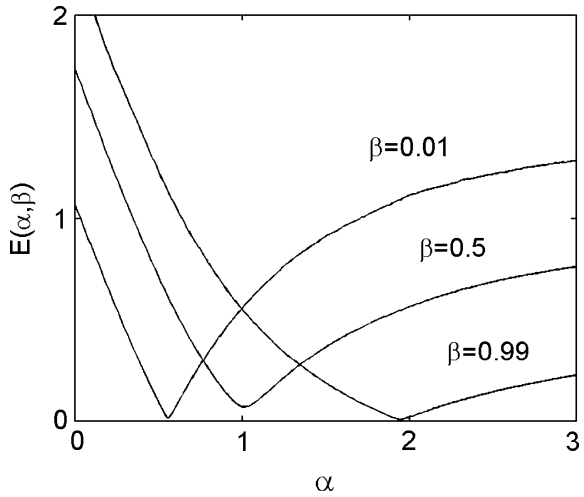

Fig. 2. Fitting error of Rényi diversities based on reassociation kinetics. The curves correspond to horizontal sections of Fig. 1 at $\beta = 0.01$, $\beta = 0.5$ and $\beta = 0.99$. The fitting error for $\beta = 0.99$ reaches a minimum close to $\alpha = 2$, showing that $\tau_{0.99}$ is strongly correlated with the Simpson diversity. Similarly, the error for $\beta = 0.5$ has a minimum close to $\alpha = 1$, and thus contains the Shannon diversity. The error $\beta = 0.01$ shows the accuracy of the best estimate for Rényi diversities with small α, like the logarithmic richness ln S.

communities. In that case, $n_s = 1/S$ for all s, $R_\alpha = \ln S$ for all α, and

$$c(\tau) = \frac{1}{\left(1 + \frac{\tau}{S}\right)^{\gamma}} \quad \text{such that} \quad \tau_\beta = S\left(\beta^{-\frac{1}{\gamma}} - 1\right)$$

Therefore, the relation between R_α and $\ln \tau_\beta$ is given by

$$\ln\tau_\beta = R_\alpha + \ln\left(\beta^{-\frac{1}{\gamma}} - 1\right) \qquad (2)$$

For the Shannon and Simpson index, this straight line with slope 1 coincides with the simulation data (Fig. 3). Interestingly, it yields a lower bound for the number of species S, an argument already used by Torsvik et al. (1990a).

Discussion

By simulating microbial communities and generating reassociation curves, we have shown that these curves encode a range of Rényi diversities. The Simpson diversity seems to be most easily accessible (for denaturation fractions $\beta \approx 1$). For longer experiments, also accurate estimates of the Shannon diversity are obtained (at $\beta \approx 0.5$), together with intermediate Rényi diversities R_α for α between 1 and 2. If the experiment can be continued until most of the DNA is reassociated (corresponding to small β), then estimation of Rényi diversities R_α with α as small as 0.5 is possible.

Although the reassociation model is highly idealised, it contains crucial information about the diversity estimation problem. For pairs (α, β) on the minimal error curve (Fig. 1), the correlation between R_α and $\ln \tau_\beta$ is almost perfect. This implies that the reassociation times τ_β with $\beta \in]0,1[$ can be mapped directly to the Rényi diversities R_α with $\alpha \in]0.5,2[$. In other words, Cot curves encode accurate information about this range of R_α and nothing else. The problem of estimating other diversity measures based on reassociation kinetics therefore reduces to the problem of determining these measures from the given range of Rényi diversities.

Looking at the estimation of the number of species S, we have to extrapolate the segment R_α with $\alpha \in]0.5,2[$ towards $\alpha = 0$. This extrapolation problem as such implies a significant loss of precision. Alternatively, one could make assumptions on the community structure (lognormal SAD, for example), but as our knowledge about microbial SADs is very limited, this introduces an important factor of arbitrariness. On the other hand, we can easily get a lower bound for the richness S, as the Rényi diversities R_α are decreasing in α. This seems to be the only reliable information available about the number of species S.

Adding more complexity to the reassociation model (like repeated sequences or species-dependent parameters) will lead to similar conclusions. The estimation of the Simpson and Shannon diversities can be expected to become less accurate, and estimating the number of species S will be even more problematic. One way to handle this situation is a careful error analysis of the richness estimation (Gans et al. 2005, Bunge et al. 2006, Volkov et al. 2006). We believe, on the contrary, that working with appropriate

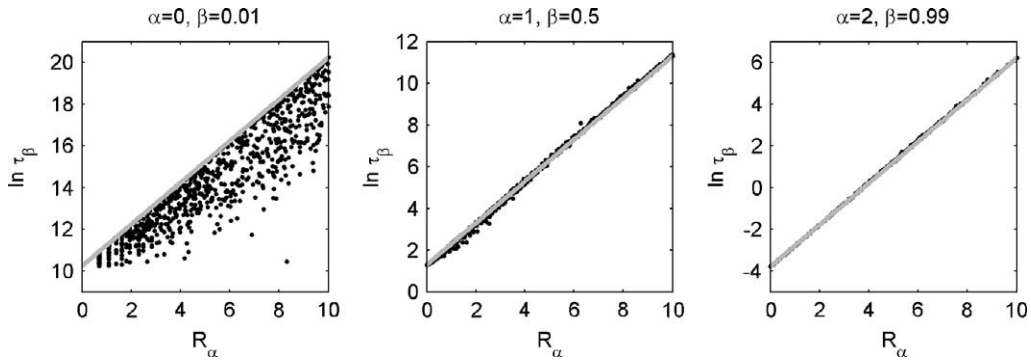

Fig. 3. Correlation between Rényi diversity R_α and reassociation time τ_β, and its analytical approximation (Eq. 2). (a) reassociation time $\ln \tau_{0.01}$ vs logarithmic richness. (b) reassociation time $\ln \tau_{0.5}$ vs Shannon diversity. (c) reassociation time $\ln \tau_{0.99}$ vs Simpson diversity. Equation 2 gives a lower bound in the first case, and an excellent approximation for cases (b) and (c).

diversity indices might be more helpful. Here we have shown that, in the case of reassociation kinetics, the data itself suggests what index is appropriate. The family of Rényi diversities can be considered as a framework to translate experimental results in terms of community structure. Estimating other diversity indicators, such as species richness, is then only possible at the expense of estimation precision.

Our analysis shows the remarkable power of Rényi diversities (or equivalently, Hill numbers) to investigate the problem at hand. Although these concepts were introduced in ecology several decades ago, practical applications have remained scarce. On the other hand, the problems in estimating microbial diversity are considerable. As standard techniques (like non-parametric richness estimators, or estimates based on SAD assumptions), often borrowed from studies of macro-organisms, seem to be of limited utility, microbial ecology is calling for fresh approaches. This contribution demonstrates how Rényi diversities can help to fill this gap.

References

Britten, R. J. and Kohne, D. E. 1968. Repeated sequences in DNA. – Science 161: 529–540.

Bunge, J. et al. 2006. Comment on ''Computational improvements reveal great bacterial diversity and high metal toxicity in soil''. – Science 313: 918c.

Choi, I. G. and Kim, S. H. 2007. Global extent of horizontal gene transfer. – Proc. Natl Acad. Sci. USA 104: 4489–4494.

Curtis, T. P. et al. 2002. Estimating prokaryotic diversity and its limits. – Proc. Natl Acad. Sci. USA 99: 10494–10499.

Forney, L. J. et al. 2004. Molecular microbial ecology: land of the one-eyed king. – Curr. Opin. Microbiol. 7: 210–220.

Gans, J. et al. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. – Science 309: 1387–1390.

Gevers, D. et al. 2005. Reevaluating prokaryotic species. – Nat. Rev. Microbiol. 3: 733–739.

Godde, J. S. and Bickerton, A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. – J. Mol. Evol. 62: 718–729.

Hill, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. – Ecology 54: 427–432.

Hong, S. H. et al. 2006. Predicting microbial species richness. – Proc. Natl Acad. Sci. USA 103: 117–122.

Lande, R. et al. 2003. Stochastic population dynamics in ecology and conservation. – Oxford Univ. Press.

Loisel, P. et al. 2006. Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revised by simulation and used as a tool to measure microbial diversity. – Environ. Microbiol. 8: 720–731.

Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.

Murugan, R. 2003. Revised theory on DNA renaturation kinetics and its experimental verification. – Biochem. Biophys. Res. Comm. 293: 870–873.

Rényi, A. 1961. On measures of entropy and information. – In: Neyman, J. (ed.), Proc. 4th Berkeley Symp. Math. Stat. Probabil. Univ. of California Press, pp. 547–561.

Schloss, P. D. and Handelsman, J. 2006. Toward a census of bacteria in soil. – PLoS Comp. Biol. 2: e92.

Stackebrandt, E. et al. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. – Int. J. Syst. Evol. Microbiol. 52: 1043–1047.

Torsvik, V. et al. 1990a. Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. – Appl. Environ. Microbiol. 56: 776–781.

Torsvik, V. et al. 1990b. High diversity in DNA of soil bacteria. – Appl. Environ. Microbiol. 56: 782–787.

Venter, J. C. et al. 2004. Environmental genome shotgun sequencing of the Sargasso sea. – Science 304: 66–74.

Volkov, I. et al. 2006. Comment on ''Computational improvements reveal great bacterial diversity and high metal toxicity in soil''. – Science 313: 918a.