OIKOS 114: 479–493, 2006

A statistical approach to estimate soil ciliate diversity and distribution based on data from five continents

Anne Chao, P. C. Li, S. Agatha and W. Foissner

Chao, A., Li, P. C., Agatha, S. and Foissner, W. 2006. A statistical approach to estimate soil ciliate diversity and distribution based on data from five continents. – Oikos 114: 479–493.

A total of 359 soil samples collected from five continents (Africa, Asia, Australia, Europe and South America) were investigated for the presence/absence of soil ciliate species. Merging records by species identity, we have compiled a master data list (species by sample matrix). In the list, a total of 964 soil ciliate species (644 described and 320 undescribed) are recorded. The species distributions within the 359 samples and across the five continents are examined. The frequency distribution of the species over samples is used for global diversity estimation. A statistical ACE (abundance-based coverage estimation) model which links observed data to unseen species is applied to assess regional and global soil ciliate species richness. The model, whose reliability was tested by its power to predict the number of new species in additional samples from Africa, may resolve the controversial issue on global species diversity of soil ciliates. Although an accurate point estimate is not feasible due to severe undersampling, the statistical model enables us to obtain a minimum regional diversity and global species diversity. A consistent finding over all five continents is that at least half of the species diversity is still undiscovered. Our model also yields a global soil ciliate diversity of at least 1900 species, that is, doubles the number of currently known species, and thus diversity is relatively high. This is consistent with the finding of Foissner, who used a probability-based method. Soil ciliate distributions between continent pairs are analyzed by adjusted abundance-based similarity/overlap indices. These new indices account for the effect of unseen species and also reduce the bias generated by undersampling. The adjusted abundance-based Jaccard (or Sørensen) index shows that there is about 30% (18% for Sørensen) dis-similarity between any two continents, supporting the moderate endemicity model. The results are discussed with respect to protist species distribution, that is, whether they are cosmopolitan or of restricted distribution.

A. Chao and P.C. Li, Inst. of Statistics, National Tsing Hua Univ., Hsin-Chu, 30043 Taiwan. S. Agatha and W. Foissner, FB Organismische Biologie, Univ. Salzburg, Hellbrunnerstrasse 34, AT-5020 Salzburg, Austria (wilhelm.foissner@sbg.ac.at).

For many biological communities, a complete census of species is almost unattainable. It has long been recognized that there are unseen species in almost every taxonomic survey or species inventory. In the same way that surveys are used to predict election results by a small sample of data from a population, biologists often estimate diversity based only on data taken from the target community. Statistical models play an important role of linking data (“seen” species) and community (which includes “unseen” species). For estimating species richness in a community, a variety of statistical models have been proposed; see Bunge and Fitzpatrick (1993),

Accepted 31 January 2006
Subject Editor: Heikki Setälä

ISSN 0030-1299

Colwell and Coddington (1994) and Magurran (2004) for comprehensive reviews, and Chao (2005) for a brief summary.

A natural approach is to extrapolate a species–accumulation curve or species–area curve to predict its highest point. Flather (1996), Gotelli and Colwell (2001) and Magurran (2004) provide reviews of various curves and their associated species patterns. Statistical sampling-based approaches include parametric and nonparametric methods. The parametric method uses a parametric distribution to fit the observed frequency counts. A classic example is the gamma-type distribution for species abundance as proposed by Fisher et al. (1943). Preston's log-normal curve and MacArthur's broken-stick model are further examples. The nonparametric approach generally utilizes frequencies of rare species to estimate the number of unseen species, because rare species carry almost all information about the missing species.

The three approaches mentioned above have been widely applied to estimate species richness for macroorganisms. A popular computer program, EstimateS, developed by Colwell (1994–2004), is available. Chao and Shen (2003–2005) also developed a program, SPADE, featuring various estimators. Very recently, microbiologists have discovered that statistical models may be used to count the uncountable microbial diversity (Hughes et al. 2001). In a minireview, Bohannan and Hughes (2003) introduced parametric and nonparametric methods as new approaches for analyzing microbial biodiversity data. Stach et al. (2003) applied statistical models to estimate actinobacterial diversity based on 16S rDNA clone libraries.

In protists, Fenchel (1993) and Finlay (2002) suggest a low diversity because the small size and high abundance of micro-organisms favour global dispersal and thus low rates of allopatric speciation. They explain the lack of certain micro-organisms in certain areas as a result of uneven sampling efforts. In contrast to this "cosmopolitan model", the "moderate endemicity model", mainly put forward by Foissner (1999a, 2004b, 2005, 2006) and Foissner et al. (2002), refers to the many eyecatching "flagship species" which have never been found in other well investigated areas, and emphasizes that most protists are much older than multicellular organisms and, thus, had sufficient time to acquire considerable diversity. This is supported by the continuous discovery of new flagship species showing our ignorance about even conspicuous taxa (Foissner 2004a, 2005, Fig. 1–8). Further, the restricted distribution of macrofungi and mosses, which disperse by spores much smaller than most protists, shows that minuteness and high abundance do not necessarily imply global distribution (Foissner 2006). For instance, a single *Agaricus campestris* (mushroom) releases 1.6×10^{10} spores in six days (Webster 1983), which exceeds the abundance of ciliates in a m^2 of forest soil by several orders of magnitude (Meyer et al. 1989).

Our investigation has two goals. First, we estimate minimum regional and global numbers of soil ciliates, based on an investigation of 359 soil samples from five continents using new statistical tools. Second, we calculate an adjusted abundance-based similarity index between any two continents to support the moderate endemicity model.

Material and methods

Regions and sampling

A total of 359 samples were collected from five continents (82 samples from Africa, 29 from Asia, 164 from Australia, 34 from Europe and 50 from South America) to record the presence/absence of soil ciliate species. There were two sub-region records in Africa, Europe and South America (Table 1). Generally, collections were made from a variety of biotopes covering most principal soil and vegetation types of the respective region. Detailed site descriptions are available for about half the samples (Blatterer and Foissner 1988, Berger and Foissner 1989, Foissner 1995, 1997a, 1998, 1999b, 2000, 2005, Foissner et al. 2002, 2005), while the others will be provided in later publications.

The material collected usually included mineral top soil (0–5 cm depth) with fine plant roots, the humic layer, and the deciduous and/or grass litter from the soil surface. Furthermore, many samples contained some terrestrial or aboricolous mosses with adhering soil and/or bark from trees. All these habitats are referred to as "terrestrial", as opposed to freshwater, because they contain, although in varying amounts, true humic and mineral soil (Foissner 1987). Usually, about 10 small subsamples were taken from an area of about 100 m^2 and mixed to a composite sample. All samples were air-dried for at least one month and then sealed in plastic bags.

Sample processing and community analysis

All collections were analysed with the non-flooded petri dish method as described in Foissner (1987) and Foissner et al. (2002). Briefly, this simple method involves placing 10–500 g terrestrial material in a petri dish (10–15 cm in diameter) and saturating but not flooding it with distilled water. Such cultures were analysed for ciliates by inspecting about 2 ml of the run-off on days 2, 7, 14, 21 and 28. The non-flooded petri dish method is selective, that is, only about one third of the species, described and undescribed, actually present in a single sample can be reactivated from the resting cysts (Foissner 1997a, 1999a, Foissner et al. 2002, 2005).

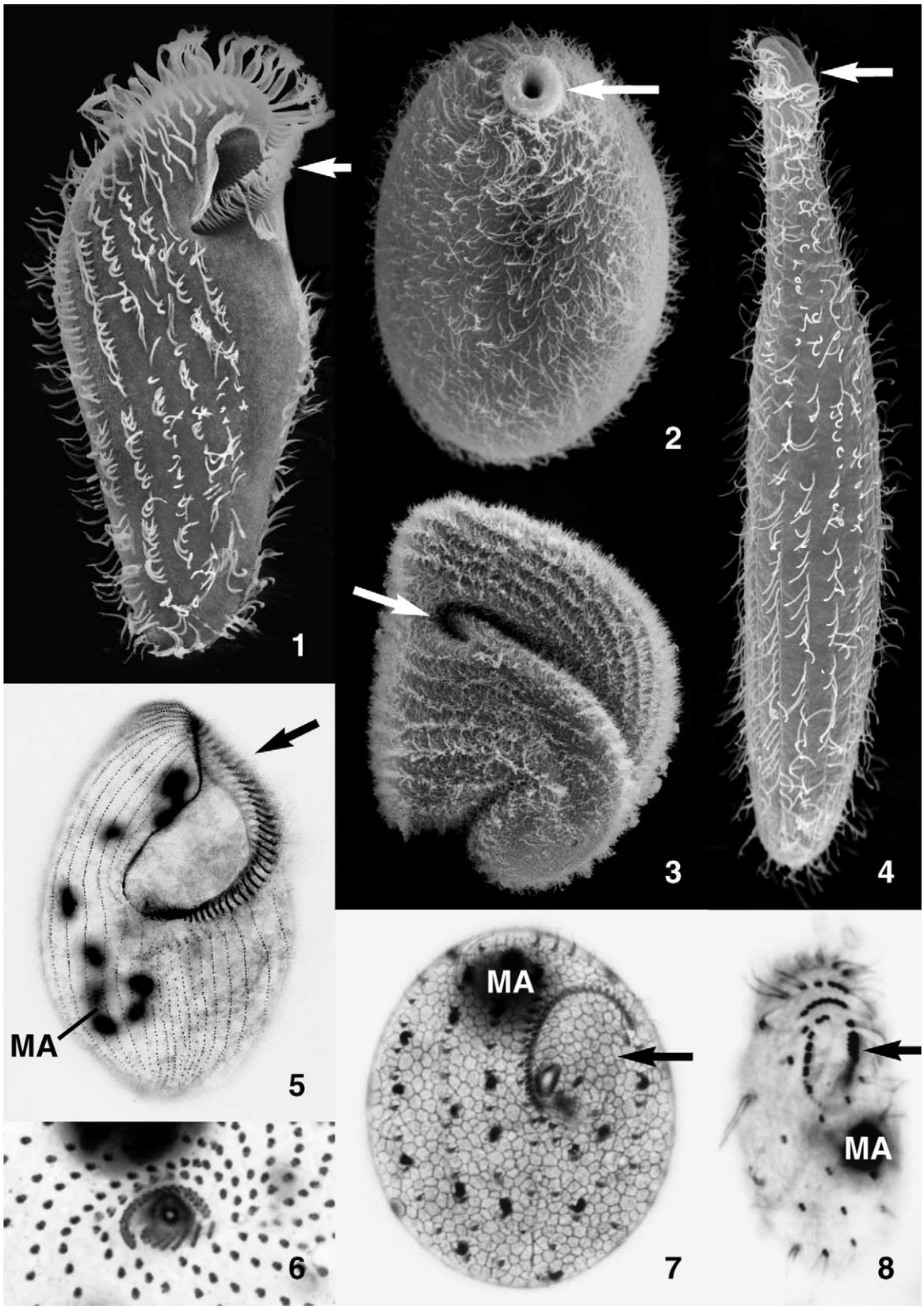

Fig. 1-8 (*Continued*)

Table 1. Data summary for soil ciliates from five continents.

Region	Subregion	Number of samples	Data sets[1]			Shared species in two subregions
			a	b[2]	c[9]	
Africa		82	399	58	75	91
	Kenya[4]	9	125	0	12	
	Namibia[4]	73	365	58	63	
Asia[6]		29	188	46		
Australia[7]		164	314	102	69	
Europe		34	352	35	16	154
	Austria[5]	14	245	30	12	
	Germany[3,5]	20	261	5	4	
South America[8]		50	246	79		102
	Costa Rica[6]	34	208	49		
	Amazon[6]	16	140	30		
Total (global)		359	644	320		

[1]a – described species in the master list, b – undescribed species in the master list, c – species not in the master list but, included in regional analyses (see "data acquisition and treatment" for details)
[2]Those species are tentatively labelled as "undescribed 1", "undescribed 2..." in the master list
[3]Several of the German sites are composites of up to 298 sub-sites. However, taxonomy of soil ciliates was still in its infancy when these sites were investigated. Thus, many undescribed species were possibly overlooked or even misidentified
[4]Published in Foissner (1999b) and Foissner et al. (2002)
[5]Published in Foissner (2000) and Foissner et al. (2005)
[6]Only partially published (Berger and Foissner 1989, Foissner 1993, 1995, 1997a, Foissner et al. 2002, Foissner and Xu 2005)
[7]Only partially published (Blatterer and Foissner 1988, Foissner 1993, Foissner and Xu 2005, Foissner et al. 2002)
[8]Reliable numbers are lacking for set (c) because the detailed investigation of the preparations is still in progress. This only affects regional analysis (Table 4)
[9]Frequencies assumed: These classifications are based on data of Foissner (1998)
Kenya: 2 species occurring in three samples, 2 in two samples, and 8 in only one sample;
Namibia: 3 species occurring in four samples, 10 in three samples, 20 in two samples, and 30 in only one sample;
Australia: 4 species occurring in three samples, 15 in two samples, and 50 in only one sample;
Austria: 2 occurring in three samples, 4 in two samples, and 6 in only one sample; Germany: all 4 occurring in only one sample

Unfortunately, a better method is not known and a combined morphological/molecular approach for ciliate identification is still impossible for larger collections because (i) single specimen PCR is still a problem in ciliates; (ii) a single soil sample contains an average of 40 ciliate species, making molecular identification very time-consuming (for the present study, 15 000 single PCRs would have been necessary; 75 000 if considered that each sample has been investigated five times); and (iii) a reliable method for isolation of ciliate DNA from raw soil samples has not yet been described.

Identification of species and species concept

Identification of species was according to the literature cited in Foissner (1998) and Foissner et al. (2002). Most of the species found were either new or described (or redescribed) by Foissner and students/collaborators. Thus, identification was mainly of live specimens using a high-power oil immersion objective and differential interference contrast. However, all "difficult", new or supposedly new species were investigated with the silver staining techniques described in Foissner (1991).

Fig. 1-8. Examples of soil ciliate flagship species with, likely, restricted geographic distribution. Both, scanning electron microscopy (1–4) and silver impregnation (5–8) were used for the identification of the ciliates in the present study. These methods reveal finest features and are thus indispensable in modern ciliate taxonomy. Arrows mark mouth area. MA – macronucleus (1) So far, *Saudithrix terricola*, an about 270 μm long, highly characteristic stichotrich ciliate, has been found only in field soil from Saudi Arabia (from Berger et al. 2006). (2) *Enchelydium blattereri* was discovered in floodplain soil from Australia. This conspicuous species, which belongs to the haptorid gymnostomes, has a length of about 240 μm and a highly characteristic oral bulge (from Foissner et al. 2002). (3) This is a not yet described colpodid flagship from a green river bed in Botswana, Africa. It has a length of about 300 μm and is distinctly spiralized. (4) A not yet described, about 200 μm long *Spathidium* (haptorid gymnostome) from soil of the Galapagos Islands. (5) A not yet described, about 250 μm long heterotrich ciliate from soil of a mangrove forest in Venezuela. This species, which belongs to the genus *Condylostomides*, is a flagship because it is large and green due to countless cortical granules. (6) *Fungiphrya strobli* is a functional flagship, that is, belongs to the obligate mycophagous colpodids. So far, this species has been found only in soil from the Table Mountain in Cape Town, Republic of South Africa. The unique oral apparatus is recognizable in the centre of the micrograph. There are a semicircular undulating membrane on the upper margin of the oral area and eight short adoral ciliary rows on the lower margin of the oral area. Between undulating membrane and adoral ciliary rows, there is a black circle with a bright centre, that is, the about 2 μm long feeding tube, used to penetrate fungal hyphae and to transport their contents into the ciliate (from Foissner 1999c). (7) *Apocolpodidium (Phagoon) macrostoma* is only 50 μm long, but conspicuous due to the huge oral apparatus with a semicircular undulating membrane. As yet, this species has been found only in soil from the Everglades of Florida, USA (from Foissner et al. 2002). (8) *Pseudokreyella etoschensis* was discovered in the Etosha Pan, Namibia. Although it is only 20 μm long, it is a morphological flagship due to the complex somatic and oral ciliary pattern (from Foissner et al. 2002).

Usually, these methods yield permanent slides which have been or will be deposited in the Oberösterreichische Landesmuseum in Linz (LI).

The species concept, of course, influences the number of species found and/or recognised as "undescribed". We usually apply the morphospecies concept, being aware that it considerably underestimates diversity because taxonomic resolution is coarser for protists than for plants and animals at the present state of art. This has been shown by genetic and molecular studies (Dini and Nyberg 1993, Nanney et al. 1998, Katz et al. 2005, Scheckenbach et al. 2005). Thus and for the reasons mentioned in the foregoing section, our species numbers are, on average, likely underestimated by an order of magnitude, both as concerns described and undescribed (new) species.

Data acquisition and treatment

Our multiple-sample data were merged by species identity to form a master data list (available from the authors in Excel format). We deal with both regional and global diversity estimation, but only part of the taxonomic descriptions have been published (Table 1). Thus, we had to adjust statistical analyses to suit the data available. In our data, we distinguished three data subsets, (a), (b) and (c), as discussed below. Only subsets (a) and (b) were listed in the merged Excel sheet. Data subsets (c), used only for regional analysis, were not included in the list.

Data set (a): "described species", that is, species with name in the master list. Although the "cutting years" (i.e. years used to classify species into described or undescribed) are different in the original source data sets in each subregion or region listed in Table 1, we unified them based on the species classification/status in our list. For example, Foissner et al. (2002) identified 365 ciliate species in 73 soil samples from Namibia. Of these, 128 were undescribed (new) species before 2002. These new species were then described for the first time in the 2002 monograph. Using a cutting year 2001, these 128 species would be classified as "undescribed". However, because these 128 species have been given names since 2002 and their names are recorded in our list, they are treated as described for the present calculations (Table 1). We remark that the status (described or undescribed) of a species does not affect our statistical analyses; only its frequency data are relevant.

Data set (b): "undescribed species", that is, species without names in the master list. Conservatively estimated, there are 320 such species; they are tentatively and consecutively numbered "undescribed1", "undescribed2" ..., in the list. All these species have been well studied, but the findings have not yet been published. Most of these species have been found only in the continent in which they were discovered. Thus, we assume that they are not shared by the other regions. For Namibia, we have good reasons to include 58 unidentified, and most probably undescribed, species in this category. Non-identification is mainly due to the lack of sufficient specimens, observations and/or preparations, as explained in Foissner et al. (2002). Half of these 58 species were also discovered in further 20 samples from Namibia (W. Foissner, unpubl.).

Data set (c): "Species not in the list", that is, species collected in our study but not listed in the master sheet. They are considered only in regional diversity analysis and their frequencies are given in footnote 9 of Table 1. Most of these taxa are undescribed species, but the number of individuals is too low or the preparations are too poor for a formal description. In Namibia, for instance (Foissner et al. 2002), several of these species were also found in recent investigations (W. Foissner, unpubl.; see also the discussion about the ACE approach), justifying their inclusion in at least the regional analyses.

The three data sets (a, b, c) were used to provide minimum estimates for species diversity and proportions of unseen species in each subregion or region; sets (a, b) were used to provide an minimum estimate of global soil ciliate diversity and for similarity analysis.

Models and statistics

In our statistical model, a "community" refers to those species that would be found if an infinite number of samples are taken from the study area and processed with the non-flooded petri dish method. Based on available data from only a finite number of samples, our estimating target (species diversity) is the number of species in such a community. We use the abundance-based coverage estimator (ACE) to estimate minimum regional and global numbers of soil ciliate species. The ACE approach was originally developed by Chao and Lee (1992) for moderately heterogeneous communities when data information is concentrated on low frequencies; Chao et al. (1993) subsequently extended it for long-tailed frequency data in highly heterogeneous communities. Here heterogeneity is characterized by the degree of variation among species abundances. Chao and Shen (2003–2005) and Chao (2005) provide short reviews. Both EstimateS and SPADE feature ACE estimates. In this approach, all 359 samples were pooled together as a single large sample collection from the global community. Since most samples are standardized to have the same volume of soil, the frequencies of species occurrences in our samples (Tables 2, 4) can be regarded approximately as species abundances. The method is based on frequency data of rare species to estimate the number of unseen species. Appendix A

Table 2. Species frequency distribution over 359 samples based on data sets (a) and (b) (f_m: number of species found in m samples out of all samples investigated).

(a) Described species (644 species)												
m	1	2	3	4	5	6	7	8	9	10		
f_m	166	112	51	33	32	19	18	17	20	8		
m	11–20	21–30	31–40	41–50	51–60	61–70	71–100	101–150	151–278			
f_m	61	33	17	6	9	4	16	10	12			
(b) Undescribed species (320 species)												
m	1	2	3	4	5	6	7	8	10	11	12	19
f_m	238	48	12	9	1	3	1	3	1	2	1	1

contains the ACE formulas with model assumptions. The ACE estimator and its variance depend only on species frequencies.

The incidence-based coverage estimator (ICE), for which samples are not pooled, has also been applied to our data (Chao 2005). The theory of Kendall (1999) justifies that the method can be extended to deal with multiple samples. However, the evaluation of the variance of the ICE estimator as applied to our regional diversity analysis requires detail species distribution over samples for data sets (c), thus the variances are not obtainable for any subregion/region with sets (c). For the other subregions/regions without sets (c) and for global analysis in which sets (c) were not considered, the ICE approach yields estimates very close to those from the ACE approach. Therefore, we only present the ACE results in diversity analyses.

To compare species distributions of soil ciliates between any two continents, we propose using the adjusted abundance-based Jaccard and Sørensen similarity/overlap indices recently developed by Chao et al. (2005, 2006) and modified by Li (2005) to adapt for the case of multiple samples. The classic incidence-based Jaccard and Sørensen indices are subject to serious under-sampling biases.

Results

Species distribution over 359 samples

Based on the presence/absence of each species within the 359 samples (data sets (a) and (b) in Table 1), we obtained the frequency distribution for each species in our master data list. Table 2 gives the frequency distribution separately for 644 described and 320 undescribed species. Because described and undescribed species do not overlap, we can collapse the two tables to get the frequencies for all observed species. The first 10 frequencies for all 964 species are shown in the last row of Table 4; they will be used for global diversity estimation in subsequent sections.

For the undescribed species, we assume that each species occurred only in one continent, but it might be found in several samples within that continent. Nevertheless, there were 238 species occurring in one sample and the most frequent one occurred in 19 samples. For the described species, there are some very frequent and common species (51 species occurred in more than 50 samples, 22 species occurred in more than 100 samples, 12 species occurred in more than 150 samples, and the most frequent species occurred in 278 samples), but there is a considerable proportion (43%) of infrequent species (166 in one sample and 112 in two samples, out of a total of 359 samples). The data include both low and high frequencies, implying a high degree of heterogeneity for the soil ciliates community.

Species distribution over five continents

The 964 species in our master list can be classified into 31 categories according to species presence/absence in each of the four continents. For notational convenience, we denote 0 for absence of a species and 1 for presence of any species. Then species presence-absence data can be categorized as shown in Table 3. For example, the first category (0, 0, 0, 0, 1) indicates that there are 105 species (26 described) found only in South America; the second (0, 0, 0, 1, 0), fourth (0, 0, 1, 0, 0), eighth (0, 1, 0, 0, 0) and sixteenth categories indicate that, respectively, there are 137, 134, 48 and 191 species found only in Europe, Australia, Asia and Africa. There are a total of $105+137+134+48+191=615$ species (64%) occurring only in one continent. Those 615 species represent the "observed" unique species. This high proportion is, of course, obtained because we assumed that all undescribed species are found only in one continent. However, if we exclude all undescribed species and only consider those described in Table 3, there is still a high percentage (46%) of species that occurred only in one continent. This percentage is consistent with Foissner (1998), who found a corresponding rate of 44%. Based on Table 3, the percentages of observed unique species divided by the observed regional total range from 21%–36%, gives an observed average rate of 30%. To compare species distributions, we assess similarity/overlap indices between any two continents. Relevant comparisons are presented after species diversity analysis.

Table 3. Species distribution according to presence/absence in each of the five continents (0 denotes absence and 1 denote presence).

Category number	Africa	Asia	Australia	Europe	South America	Described species and undescribed	Described species only
1	0	0	0	0	1	105	26
2	0	0	0	1	0	137	102
3	0	0	0	1	1	7	7
4	0	0	1	0	0	134	32
5	0	0	1	0	1	13	13
6	0	0	1	1	0	14	14
7	0	0	1	1	1	15	15
8	0	1	0	0	0	48	2
9	0	1	0	0	1	2	2
10	0	1	0	1	0	8	8
11	0	1	0	1	1	1	1
12	0	1	1	0	0	3	3
13	0	1	1	0	1	6	6
14	0	1	1	1	0	6	6
15	0	1	1	1	1	8	8
16	1	0	0	0	0	191	133
17	1	0	0	0	1	10	10
18	1	0	0	1	0	18	18
19	1	0	0	1	1	5	5
20	1	0	1	0	0	28	28
21	1	0	1	0	1	14	14
22	1	0	1	1	0	20	20
23	1	0	1	1	1	19	19
24	1	1	0	0	0	9	9
25	1	1	0	0	1	0	0
26	1	1	0	1	0	3	3
27	1	1	0	1	1	4	4
28	1	1	1	0	0	5	5
29	1	1	1	0	1	9	9
30	1	1	1	1	0	15	15
31	1	1	1	1	1	107	107

Table 3 shows that there are 107 species (12%, the last row in Table 3) found in all five continents. The existence of both frequently and infrequently occurring species again shows a high degree of heterogeneity. Foissner (1998) compiled a world soil ciliate list over five geographical regions based on 817 samples, but species presence/absence for each sample was not recorded. Thus, regional species abundances are not available, so that the list neither can be used statistically for regional diversity estimation, nor for a global estimate without further restrictive assumptions. See Discussion for a description of an updated list by combining the data sets (a, b) in Table 1 and those in Foissner (1998).

Estimates for regional and global soil ciliate diversity

Using the ACE method and the program SPADE with the data sets (a, b, c) in Table 1, we obtained minimum ciliate species diversity within each subregion and region (Table 4). To take account of sampling variation, we

Table 4. Regional and global soil ciliate species diversity.

Region	Subregion	Estimate of minimum species diversity	95% confidence interval	Percentage of unseen species	First 10 frequencies ($f_1 \sim f_{10}$) used in calculations
Africa		900	(757, 1134)	41%	(221, 104, 54, 36, 22, 11, 14, 7, 6, 4)
	Kenya	217	(180, 285)	37%	(60, 21, 24, 12, 9, 3, 3, 2, 3, 0)
	Namibia	830	(685, 1078)	41%	(204, 98, 45, 38, 17, 13, 11, 7, 6, 4)
Asia		463	(319, 844)	50%	(98, 51, 16, 6, 5, 10, 6, 5, 2, 8)
Australia[1]		865	(703, 1148)	44%	(190, 95, 36, 23, 13, 15, 12, 9, 5, 8)
Europe		706	(575, 937)	43%	(159, 72, 27, 25, 19, 7, 9, 14, 7, 5)
	Austria	446	(367, 601)	36%	(102, 49, 26, 20, 12, 6, 13, 5, 11, 5)
	Germany	552	(384, 971)	51%	(124, 56, 21, 18, 6, 6, 6, 8, 4, 6)
South America[1]		638	(477, 971)	49%	(145, 64, 20, 23, 11, 5, 11, 5, 7, 3)
	Costa Rica	576	(377, 1100)	55%	(126, 39, 23, 14, 10, 7, 6, 3, 4, 4)
	Amazon	317	(254, 426)	46%	(79, 31, 21, 12, 7, 4, 3, 3, 2, 2)
Global[2]		1928	(1600, 2427)	50%	(404, 160, 63, 42, 33, 22, 19, 20, 20, 9)

[1] Data set (c) is not available; inclusion of set (c) in the analysis will generally result in a higher regional estimate
[2] Based only on data sets (a, b) for conservative purposes

present the 95% confidence interval. The estimated percentage of unseen species in each subregion/region is also given. Because the first 10 frequencies are the data used for estimating the number of unseen species (Appendix A), they are listed in the last column. From the estimated minimum species richness, it is statistically difficult to determine the true species richness, but a consistent finding over all subregions/regions is that at least half of the species are not seen in our samples. To be conservative, the global diversity is computed based on data sets (a, b) only. Our model concludes that global soil ciliate diversity is at least 1900 with a 95% interval of (1600, 2400). Thus, global soil ciliate species richness is at least double the number currently identified (Foissner 1998, Foissner et al. 2002). The ICE approach gives a very close estimate.

We remark that any estimate refers to the minimum species diversity of the community in the study area. For example, an estimate of 900 species for Africa refers to the species richness of the combined sampling areas of Kenya and Namibia. Thus, we cannot interpret this estimate to include species in Tanzania because no data were taken from there. The value of 865 species for Australia refers to the diversity of the community in the study area there. Similar interpretations pertain to the other regions/continents. See Discussion for further statistical reasons that our estimates represent minimum species richness.

Comparing soil ciliate distributions

Table 5 shows the classic Jaccard and Sørensen similarity indices based on data sets (a, b) for ten pairs of continents. The similarity index for comparing two identical communities is always equal to one. The similarity matrix is symmetric, so we present only its lower triangle.

The classic Jaccard index is the observed fraction of shared species in a pooled region of two continents; and the Sørensen index is the shared species divided by the average number of species. Thus, the Jaccard dissimilarity (1-Jaccard index) measures the non-shared fraction in a pooled region; and the Sørensen dissimilarity (1-Sørensen index) approximately measures the average of the non-shared fraction in a region. All indices can be calculated from the counts in Table 3.

The classic Jaccard similarity indices between any pair of continents are quite low, with an average of 0.31 (dissimilarity = 0.69). The average for the Sørensen similarity index is 0.48 (dissimilarity = 0.52). However, these indices have severe biases and are sensitive to sample sizes. For example, in the Discussion, we show that the averaged Jaccard similarity index is increased to 0.36 for the updated list of world soil ciliates.

The abundance-based indices are conceptually different from the classic incidence-based ones. Thus, their values are not comparable. A simple, artificial example in the Discussion will highlight the difference between incidence-based and abundance-based indices. The incidence-based approach treats all species equally whereas the abundance-based approach treats all individuals equally. In our multiple-sample data, an individual refers to a species presence/occurrence in any soil sample. Chao et al. (2005, 2006) developed abundance-based Jaccard and Sørensen similarity indices and proposed using adjusted forms that reduce biases generated by unseen species due to undersampling. Recently, the adjusted forms are further modified by Li (2005) to adapt for the case of multiple samples. Table 6 shows the adjusted Jaccard and Sørensen similarity indices for ten pairs of continents based on data sets (a, b).

The adjusted Jaccard-type similarity indices are between 66% and 77%, with an average of 70% (dissimilarity = 30%; Table 6). The Sørensen-type similarity indices are between 79% and 87%, with an average of 82% (dissimilarity = 18%; Table 6). If data are a representative sample, we can roughly interpret this as follows: in the community, the fraction of occurrences that are classified as non-shared species is about 18% in one continent and 30% in a pooled region of two continents. We can further construct a 95% confidence interval by assessing sampling errors. Based on the bootstrap method developed by Chao et al. (2006), the

Table 5. Classic incidence-based similarity indices.

Region	Asia	Africa	Australia	Europe	South America
(a) Classic Jaccard					
Asia	1				
Africa	0.28	1			
Australia	0.32	0.33	1		
Europe	0.32	0.29	0.34	1	
South America	0.32	0.27	0.35	0.30	1
(b) Classic Sørensen					
Region	Asia	Africa	Australia	Europe	South America
Asia	1				
Africa	0.44	1			
Australia	0.49	0.5	1		
Europe	0.49	0.45	0.51	1	
South America	0.49	0.43	0.52	0.47	1

Table 6. The Chao et al. (2005) adjusted abundance-based similarity indices.

Region	Asia	Africa	Australia	Europe	South America
(a) Jaccard-type index					
Asia	1				
Africa	0.66	1			
Australia	0.74	0.73	1		
Europe	0.68	0.66	0.75	1	
South America	0.70	0.66	0.77	0.70	1
(b) Sørensen-type index					
Region	Asia	Africa	Australia	Europe	South America
Asia	1				
Africa	0.79	1			
Australia	0.85	0.84	1		
Europe	0.81	0.79	0.85	1	
South America	0.83	0.79	0.87	0.82	1

community-level Sørensen dissimilarity is in the range of 18% $\pm$ 4% (i.e. the estimated standard error is about 2%), and the Jaccard dis-similarity is 30% $\pm$ 6% (i.e. the estimated standard error is about 3%). Incorporating sampling errors, we find that the lowest community-level Sørensen dissimilarity is unlikely below 14%, and the corresponding Jaccard dissimilarity is unlikely below 24%. Consequently, our current data support the moderate endemicity model.

An updated world list of soil ciliates

The data sets (a, b) described in Table 1 have been merged with that listed in Foissner (1998) to form an updated compilation of world soil ciliates. This updated list contains 816 well-described and 320 undescribed species across five biogeographical regions. The five regions are Holarctic (including Asia and Europe; 707 species), Paleotropis (including Africa; 511 species), Australis (Australia and Tasmania; 468 species), Neotropis (including South America; 409 species) and Archinotis (including Antarctic; 101 species). Since the data for Asia and Europe are not separately available in Foissner (1998), our updated list combines them in a Holarctic region.

The master list mentioned in Table 1 contains a species by sample matrix, while the updated list contains only a species by region incidence matrix, because the detailed data for each sample are not provided in Foissner (1998). Thus, only classic, incidence-based indices can be used (Table 7). The average of the Jaccard (or Sørensen) indices is increased to 0.36 (0.56 for Sørensen), compared with 0.31 (0.48 for Sørensen) in Table 5, showing the sensitivity of the classic indices. Based on the Jaccard dissimilarity index (as distance), we plot a cluster tree, using a proper statistical linkage method (Ward's error sum of squares method; Fig. 9). The cluster clearly shows the separation of Laurasia and Gondwana, and thus provides statistical evidences for the influence of historical events on the distribution of micro-organisms.

Discussion

The ACE approach

The basic idea in the ACE approach is that abundant (or frequent) species carry almost no information regarding the number of unseen species, because abundant species would be seen by any sampler; only rare (or infrequent) species can be missed and thus carry missing species information. Our model as applied to ciliate species estimation divides the observed species into two parts: abundant (species found in more than 10 samples) and rare (species found in 10 or fewer samples). We base an estimate of unseen species on the rare species and then complete the estimate by adding in the number of abundant species. Relevant formulas are provided in Appendix A. The choice of a cut-off point of 10 is based on simulation experiments and empirical evidences. Magurran and Henderson (2003) suggested a similar cut-off for fish communities. If a higher cut-off is used (i.e. more species are classified into the group of rare species), which would be justifiable for the highly heterogeneous ciliates community, then the ACE method will result in a higher species richness. This provides a

Table 7. Classic Jaccard similarity index for an updated world list of soil ciliates.

Region	Holarctic	Paleotropis	Australis	Neotropis
Holarctic	1			
Paleotropis	0.36	1		
Australis	0.34	0.41	1	
Neotropis	0.32	0.36	0.39	1

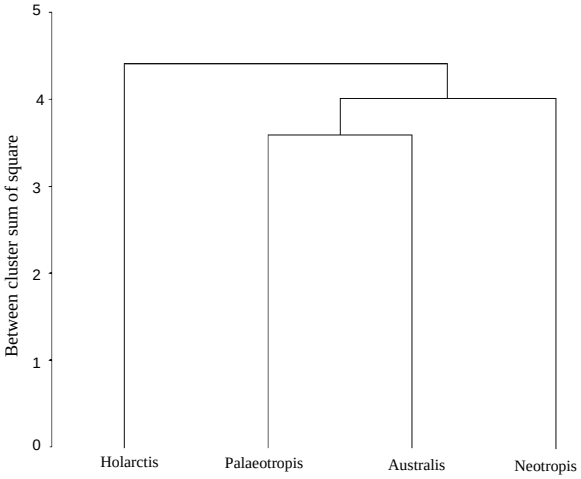

Fig. 9. World soil ciliate species cluster based on the classic Jaccard dis-similarity index as distance..

statistical reason that our diversity estimates in Table 4 represent minima. Other reasons, as discussed in the section of Material and Methods, are that the observed species numbers are minimum values due to the limits of the non-flooded petri dish method and the morphospecies concept.

From a statistical viewpoint, the undescribed species (they are generally rare species as shown in Table 1) carry most information about the number of unseen species. If we exclude those 320 undescribed and reanalyze the data based only on the 644 described species, using the frequencies in Table 2 (a), the ACE model produces an estimate of 844 with a 95% confidence interval of (780, 940). This point estimate, obtained by excluding the undescribed species, is less than half of our global estimate. Therefore, if only common described species are used, then one may conclude that species diversity is relatively low.

The data details for the Namibia region are tabulated in Foissner et al. (2002, Table 4, pp. 58-63). Foissner et al. (2002) identified 365 species and suggested, intuitively, about 1000 soil ciliate species in Namibia. Table 4 shows that our minimum for the Namibia region is 830 species with an approximate 95% confidence interval of (700, 1100), which covers 1000 as one of the plausible estimates. A test showing the reliability of our data analysis for Namibia proceeds as follows: Based on the undescribed frequencies (using a cutting year of 2001), the ACE approach gives an estimate of 360 for unseen and undescribed species in Namibia. This estimate can be used to predict the number of new species found in an additional set of five samples (Chao and Shen 2004). Relevant formulas are given in Appendix B. Our analyses predict that, if five further samples from Namibia are investigated, eight undescribed species would be found with a 95% confidence interval of (6, 9). This was, indeed, the case: in the first set of five samples, seven new species were discovered, and in a second set even eight!

Global soil ciliate diversity

Using a probability-based approach to determine a multiplier for undersampling, Foissner (1997b) estimated a global soil ciliate diversity of at least 1300-2000 species. Finlay and Fenchel (1999) argued that the multiplier is sensitive to sample size. In this paper, we used a non-parametric statistical approach for which the effect of sample size is adjusted, and we see that the results are quite consistent with the findings of Foissner (1997b).

There is also another strong indicator that the estimated 1900 species is a minimum. During the past 20 years, we found about 1000 new species of soil ciliates (for an example, see Foissner et al. 2002), and further new species are being discovered at a constant or even increasing rate, indicating that the summit is a long way off (Fig. 10).

Comparing distributions

For the analysis of distribution, similarity (overlap) or dis-similarity (distance) indices provide quantitative measures by comparing species composition in two or more habitats (Magurran 2004). A variety of similarity indices have been proposed, the two classic and most widely used ones being those of Jaccard and Sørensen. They were used, for instance, by Hillebrand et al. (2001) and Green et al. (2004) to assess protist similarity as a function of geographic distance.

The classic similarity indices, although widely used and requiring only incidence records, seriously underestimate community similarity. The biases are likely to be substantial for assemblages with high species richness and a large fraction of rare species (Wolda 1981, 1983, Magurran 2004, p.175), as is the case with soil ciliates. Chao et al. (2005, 2006) noted that they are sensitive to sample size and undersampling bias cannot easily be removed or adjusted. Hillebrand et al. (2001) and Green et al. (2004) examined similarity decrease with distance, so their conclusion of a decreasing trend are not affected by undersampling biases (because roughly the same magnitude of bias exists for all distances).

Table 8 illustrates the difference between the classic incidence-based and abundance-based indices. We consider two cases of comparison: community I vs community II and community III vs community IV. In both cases the pooled community consists of three species of which one species is shared. The classic indices are identical for these two cases: the Jaccard index is $1/3 = 0.33$ and the Sørensen is $1/2 = 0.5$. However, the species

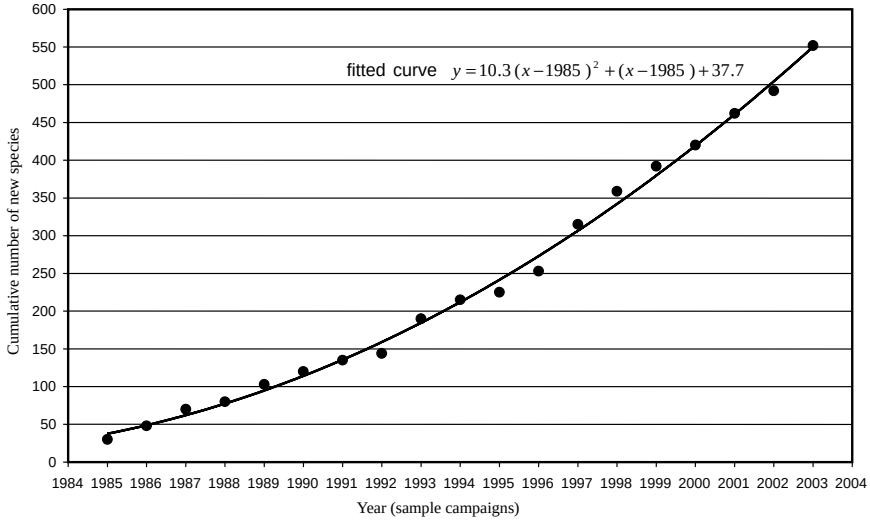

Fig. 10. Plot of the cumulative number of new ciliate species in twenty soil samples each for the past 20 years.

abundances have different distributions: in case 1 all species have the same abundance, while in case 2 one common species is shared and two rare species are not shared. The abundance-based Jaccard dissimilarity = 1- 0.73 = 0.27, and the Sørensen dissimilarity = 1- 0.85 = 0.15, implying that roughly about 15% of individuals in one community, and 27% of those in the pooled region of two communities, belong to non-shared species. This example also shows, for equal abundances, that the abundance and the incidence indices are identical. In other words, the classic approach treats all species equally and ignores their abundances.

The Chao et al. (2005, 2006) adjusted abundance-based indices, although requiring effort to collect detailed sample data, have the following properties: (i) species abundances are incorporated, and we can infer the community-level indices if the data are representative; (ii) undersampling bias can be largely removed and indices are less sensitive to sample size; (iii) sampling errors can be assessed to construct confidence intervals.

The example in Table 8 also helps to explain why the average of the classic similarity indices for soil ciliates are 0.31 (Jaccard) and 0.48 (Sørensen) in Table 5, while the adjusted abundance similarity indices jump to 0.70 (Jaccard) and 0.82 (Sørensen). As in the example, all the shared soil ciliates species are relatively common ones. The adjusted abundance-based Jaccard (or Sørensen) index shows that there is about 30% (18% for Sørensen) dissimilarity between any two continents. Seen with the biologist's eyes, a dissimilarity of about 18% - 30% is not very pronounced, but the figures, along with the estimated standard errors, are sufficient to support the moderate endemicity model (Foissner 2004a, 2004b, 2006). Foissner (2006) suggested about 30% endemic species for protists in general.

"Everything is everywhere, the environment selects"(Beijerinck 1913) is not falsifiable and thus a metaphor

Beijerinck´s statement cited above is widely considered as a scientific hypothesis. However, scientific hypotheses must be falsifiable (Popper 1962). This is not the case with Beijerinck´s statement, testing of which would

Table 8. Comparison of classic incidence-based and adjusted abundance-based similarity indices.

Communities	I	II	III	IV
	Abundance		Abundance	
Shared species a	0.5	0.5	0.9	0.8
Endemic species b	0.5		0.1	
Endemic species c		0.5		0.2
Similarity indices:				
Classic Jaccard index	0.33		0.33	
Classic Sørensen index	0.50		0.50	
Abndance-based Jaccard index	0.33		0.73	
Aundance-based Sørensen index	0.50		0.85	

require the existence of at least two identical habitats in different biogeographical regions. Further, these habitats should have a representative size and age to allow the establishment of a micro-organism community as found, for instance, in old ponds. Such conditions do not exist in the real world, and it is unlikely that they can be created experimentally. It might be important to note that "similar" or almost "identical" would be insufficient because this could implicate that differences in species composition are due to genuine differences in the habitats.

We agree with Finlay et al. (2004) that the term cosmopolitan is burdened with the same problems. A simple way to override them is to state definitely where or in which biogeographic region a certain species is present or lacking.

Not everything is everywhere: statistical, molecular and distribution evidences

Whether micro-organisms are cosmopolitan or of restricted distribution is one of the hotly discussed issues in modern microbial and protistan ecology. A "cosmopolitanism school", represented mainly by Fenchel and Finlay (2004), and a "moderate endemicity school", represented mainly by Foissner (1999a, 2004b, 2006), search for arguments to support their view, as shown in the introduction of this paper. Here, we shall briefly discuss three lines of evidences suggesting a restricted distribution of a good deal (~30%) of protist morphospecies.

The present study provides statistical evidences that support the moderate endemicity model (Foissner 2004b, 2006). Further, our data disprove the low number (about 3000) of free-living ciliate species suggested by Finlay (2001): there are at least 1900 soil ciliate species (Table 4), of which two thirds (1300) occur only in terrestrial habitats (Foissner 1998). Thus, only 1700 species remain for the limnetic and marine ecosystems, a much too low number occupied by two common groups, viz., the oligotrichs with about 1500 described species (but many possibly synonymous) and the peritrichs with about 500 described species (Corliss 1979). This view is supported by the high number of cryptic protist species revealed by molecular methods. Scheckenbach et al. (2005), for instance, investigated the SSU rDNA of five common heterotrophic flagellates from surface waters and deep-sea sediments. Three of these morphospecies contain several cryptic species. The same has been observed in the common freshwater and soil ciliate *Halteria grandinella*, where the ITS sequences revealed high genetic diversity with up to 7.6% difference between populations (Katz et al. 2005). Further examples have been reviewed in Foissner (2006).

A great deal of data show the restricted distribution of certain protist morphospecies in a wide variety of habitats, ranging from soil ciliates to benthic Antarctic foraminifera (Bonnet 1983, Dragesco and Dragesco-Kernéis 1986, Taylor and Pollingher 1987, Tyler 1996, Vyverman 1996, Foissner 1998, 2005, Meisterfeld 2000a,b, Kristiansen 2001, Pawlowski and Holzmann 2002, Foissner et al. 2003). Many of these species are "flagships" with conspicuous size, morphology, and/or colour (Fig. 1–8). They are the elephants of the microscopic world. Tyler (1996) has summarized the reasons why such taxa have the greatest probability of real endemism: "Because they are so showy, or so novel, it is unlikely that such species would be overlooked if indeed they were widely distributed. If the Australian endemics occurred in Europe or North America then they would have been seen there, long ago." Finlay et al. (2004) indicated the lack of these species in certain regions as an effect of undersampling. While this might apply to poorly studied regions, such as the tropics and remote islands, it is highly unlikely for Central Europe. Indeed, Finlay et al. (2004) never found any of these flagships in Europe. Their "European" *Nebela vas*, a famous Gondwanan testate amoebae, is based on very old, misinterpreted literature, while modern testacean biogeographers remained uncited, for instance, Bonnet (1983) and Meisterfeld (2000a,b). Likewise, Finlay et al (2002) propose that "the argument in favour of endemic diatom species is untenable because it is not possible to disprove their existence elsewhere in the biosphere". On the other hand, Finlay et al. (2002) cannot disprove the existence of endemic species with this kind of argumentation.

Small size and huge abundance are considered as the main reasons for the cosmopolitan distribution of microorganisms (Finlay 2002, Fenchel and Finlay 2004, Finlay et al. 2004). However, both arguments are flawed by (macro) fungi, mosses, and ferns. They have indisputable biogeographies (Webster 1983, Schwantes 1996), although their main dispersal means (spores) are in the size of large bacteria or small protists and are produced in gigantic numbers (Introduction). This simple fact has been ignored in the discussion of microbial distribution and diversity. Likewise, it has been ignored that protist's main dispersal means, the resting cysts, lack adaptations for air dispersal, in contrast to the seeds of many higher plants, for instance, the orchids, many of which have seeds in the size of a large *Paramecium* (~300 μm).

Acknowledgements – This study was supported by the Austrian Science Foundation, FWF projects P12367-BIO and P-15017 (for WF and SA) and the Taiwan National Science Council,

project NSC-93-2118-M007-001 (for AC and PL). The technical assistance of Dr. Eva Herzog and Mag. Birgit Peukert is greatly appreciated.

References

Beijerinck, M. W. 1913. De infusies en de ontdekking der backteriën. Jaarboek van de Koninklijke Akademie v. Wetenschappen. – Müller, Amsterdam.

Berger, H. and Foissner, W. 1989. Morphology and biometry of some soil hypotrichs (Protozoa, Ciliophora) from Europe and Japan. – Bull. Brit. Mus. Nat. Hist. (Zool.) 55: 19–46.

Berger, H., Al-Rasheid, K. A. S. and Foissner, W. 2006. Morphology and cell division of *Saudithrix terricola* n. gen., n. sp., a large, stichotrich ciliate from Saudi Arabia. – J. Eukaryot. Microbiol. (in press).

Blatterer, H. and Foissner, W. 1988. Beitrag zur terricolen Ciliatenfauna (Protozoa: Ciliophora) Australiens. – Stapfia 17: 1–84.

Bohannan, B. J. M. and Hughes, J. 2003. New approaches to analyzing microbial biodiversity data. – Curr. Opin. Microbiol. 6: 182–187.

Bonnet, L. 1983. Interet biogeographique et paleogeographique des thecamoebiens des sols. – Ann. Stn. Limnol. Besse 17: 298–334.

Bunge, J. and Fitzpatrick, M. 1993. Estimating the number of species: a review. – J. Am. Statist. Assoc. 88: 364–373.

Chao, A. 2005. Species richness estimation. – In: Balakrishnan, N., Read, C. B. and Vidakovic, B. (eds), Encyclopedia of statistical sciences, 2nd ed. Wiley (in press).

Chao, A. and Lee, S.-M. 1992. Estimating the number of classes via sample coverage. – J. Am. Statist. Assoc. 87: 210–217.

Chao, A. and Shen, T. J. 2003–2005. Program SPADE (species prediction and diversity estimation). Program and user's guide at the website http://chao.stat.nthu.edu.tw.

Chao, A. and Shen, T.-J. 2004. Non-parametric prediction in species sampling. – J. Agric. Biol. Environ. Statist. 9: 253–269.

Chao, A., Ma M. C. and Yang, M. C. K. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. – Biometrika 80: 193–201.

Chao, A., Chazdon, R. L., Colwell, R. K. et al. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. – Ecol. Lett. 8: 148–159.

Chao, A., Chazdon, R. L., Colwell, R. K. et al. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. – Biometrics 62(in press).

Colwell, R. K. 1994–2004. EstimateS: statistical estimation of species richness and shared species from samples. URL http://viceroy.eeb.uconn.edu/estimates. Persistent URL http://purl.oclc.org/estimates.

Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. – Philos. Trans. R. Soc. Lond. (B 345): 101–118.

Corliss, J. O. 1979. The ciliated Protozoa. Characterization, classification and guide to the literature. – Pergamon Press.

Dini, F. and Nyberg, D. 1993. Sex in ciliates. – Adv. Microbiol. Ecol. 13: 85–153.

Dragesco, J . and Dragesco-Kernéis, A. 1986. Ciliés libres de l'Afrique intertropicale. – Faune Tropicale 26: 1–559.

Fenchel, T. 1993. There are more small than large species? – Oikos 68: 375–378.

Fenchel, T. and Finlay, B. J. 2004. The ubiquity of small species: patterns of local and global diversity. – BioScience 54: 777–784.

Finlay, B. J. 2001. Protozoa. – Encycl. Biodiv. 4: 901–915.

Finlay, B. J. 2002. Global dispersal of free-living microbial eukaryote species. – Science 296: 1061–1063.

Finlay, B. J. and Fenchel, T. 1999. Divergent perspectives on protist species richness. – Protist 150: 229–233.

Finlay, B. J., Monaghan, E. B. and Maberly, S. C. 2002. Hypothesis: the rate and scale of dispersal of freshwater diatom species is a function of their global abundance. – Protist 153: 261–273.

Finlay, B. J., Esteban, G. F. and Fenchel, T. 2004. Protist diversity is different? – Protist 155: 15–22.

Fisher, R. A., Corbet, A. S. and Williams, C. B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. – J. Anim. Ecol. 12: 42–58.

Flather, C. H. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. – J. Biogeogr. 23: 155–168.

Foissner, W. 1987. Soil protozoa: fundamental problems, ecological significance, adaptations in ciliates and testaceans, bioindicators, and guide to the literature. – Progr. Protistol. 2: 69–212.

Foissner, W. 1991. Basic light and scanning electron microscopic methods for taxonomic studies of ciliated protozoa. – Eur. J. Protistol. 27: 313–330.

Foissner, W. 1993. Colpodea (Ciliophora). – Protozoenfauna 4: 1–798.

Foissner, W. 1995. Tropical protozoan diversity: 80 ciliate species (Protozoa, Ciliophora) in a soil sample from a tropical dry forest of Costa Rica, with descriptions of four new genera and seven new species. – Arch. Protistenk. 145: 37–79.

Foissner W. 1997a. Soil ciliates (Protozoa: Ciliophora) from evergreen rain forests of Australia, South America and Costa Rica: diversity and description of new species. – Biol. Fertil. Soils 25: 317–339.

Foissner, W. 1997b. Global soil ciliate (Protozoa, Ciliophora) diversity: a probability-based approach using large sample collections from Africa, Australia and Antarctica. – Biodivers. Conserv. 6: 1627–1638.

Foissner, W. 1998. An updated compilation of world soil ciliates (Protozoa, Ciliophora), with ecological notes, new records, and descriptions of new species. – Eur. J. Protistol. 34: 195–235.

Foissner, W. 1999a. Protist diversity: estimates of the near-imponderable. – Protist 150**:** 363–368.

Foissner, W. 1999b. Notes on the soil ciliate biota (Protozoa, Ciliophora) from the Shimba Hills in Kenya (Africa): diversity and description of three new genera and ten new species. – Biodivers. Conserv. 8: 319–389.

Foissner, W. 1999c. Description of two new, mycophagous soil ciliates (Ciliophora, Colpodea): *Fungiphrya strobli* n. g., n. sp. and *Grossglockneria ovata* n. sp. – J. Eukaryot. Microbiol. 46: 34–42.

Foissner, W. 2000. A compilation of soil and moss ciliates (Protozoa, Ciliophora) from Germany, with new records and descriptions of new and insufficiently known species. – Eur. J. Protistol. 36: 253–283.

Foissner, W. 2004a. Some new ciliates (Protozoa, Ciliophora) from an Austrian floodplain soil, including a giant, red "flagship", *Cyrtohymena (Cyrtohymenides) aspoecki* nov. subgen., nov. spec. – Denisia 13: 369–382.

Foissner, W. 2004b. Ubiquity and cosmopolitanism of protists questioned. – SILnews 43: 6–7.

Foissner, W. 2005. Two new "flagship" ciliates (Protozoa, Ciliophora) from Venezuela: *Sleighophrys pustulata* and *Luporinophrys micelae*. – Eur. J. Protistol. 41: 99–117.

Foissner, W. 2006. Biogeography and dispersal of microorganisms: a review emphasizing protists. – Acta Protozool. (in press)

Foissner, W. and Xu, K. 2005. Monograph of the Spathidiida (Ciliophora, Haptoria) volume I: Protospathidiidae, Arcuospathidiidae, Apertospathulidae. – Springer (in press).

Foissner, W., Agatha, S. and Berger, H. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa),

with emphasis on two contrasting environments, the Etosha Region and the Namib Desert. – Denisia 5: 1–1459.
Foissner, W., Strüder-Kypke, M., van der Staay, G.. W. M. et al. 2003. Endemic ciliates (Protozoa, Ciliophora) from tank bromeliads: a combined morphological, molecular, and ecological study. – Eur. J. Protistol. 39: 365–372.
Foissner, W., Berger, H., Xu, K. et al. 2005. A huge, undescribed soil ciliate (Protozoa: Ciliophora) diversity in natural forest stands of Central Europe. – Biodiv. Conserv. 14: 617–701.
Gotelli, N. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.
Green, J., Holmes, A. J., Westoby, M. et al. 2004. Spatial scaling of microbial eukaryote diversity. – Nature 432: 747–753.
Hillebrand, H., Watermann, F., Karez, R. et al. 2001. Differences in species richness patterns between unicellular and multicellular organisms. – Oecologia 126: 114–124.
Hughes, J. B., Hellmann, J. J., Ricketts, T. H. et al. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. – Appl. Environ. Microbiol. 67: 4399–4406.
Katz, L. A., McManus, G. B., Snoeyenbos-West, O. L. O. et al. 2005. Reframing the "everything is everywhere" debate: evidence for high gene flow and diversity in ciliate morphospecies. – Aquat. Microbiol. Ecol. 41: 55–65.
Kendall, W. L. 1999. Robustness of closed capture–recapture methods to violations of the closure assumptions. – Ecology 80: 2517–2525.
Kristiansen, J. 2001. Biogeography of silica-scaled chrysophytes. – Nova Hedwigia Beiheft 122: 23–39.
Li, P. C. 2005. Species estimation and similarity indices in quadrat sampling. – PhD thesis, National Tsing Hua Univ., Taiwan.
Magurran, A. E. 2004. Measuring biological diversity. – Blackwell.
Magurran, A. E. and Henderson, P. 2003. Explaining the excess of rare species in natural species abundance distributions. – Nature 422: 714–716.
Meisterfeld, R. 2000a. Order Arcellinida Kent, 1880. – In: Lee, J. J., Leedale, G. F. and Bradbury, P. (eds), An illustrated guide to the Protozoa (2nd ed.). Allen Press, pp. 827–860.
Meisterfeld, R. 2002b. Testate amoebae with filopodia.-In: Lee, J. J., Leedale, G. F. and Bradbury, P. (eds), An illustrated guide to the Protozoa (2nd ed.). Allen Press, pp. 1054–1084.
Meyer E., Foissner W. and Aescht E. 1989. Vielfalt und Leistung der Tiere im Waldboden. – Öst. Forstz. 3: 15–18.
Nanney, D. L., Park, C., Preparata, R. et al. 1998. Comparison of sequence differences in a variable 23S rRNA domain among sets of cryptic species of ciliated Protozoa. – J. Eukaryot. Microbiol. 45: 91–100.
Pawlowski, J. and Holzmann, M. 2002. Molecular phylogeny of Foraminifera – a review. – Eur. J. Protistol. 38: 1–10.
Popper, K. 1962. The logic of scientific discovery. – Harper and Row.
Scheckenbach F., Wylezich, C., Weitere, M. et al. 2005. Molecular identity of heterotrophic flagellates isolated from surface waters and deep-sea sediments of the South Atlantic based on SSU rDNA. – Aquat. Microbiol. Ecol. 38: 239–247.
Stach, J. E. M., Maldonado, L. A., Masson, D. G. et al. 2003. Statistical approaches for estimating actinobacterial diversity in marine sediments. – Appl. Environ. Microbiol. 69: 6189–6200.
Schwantes, H. O. 1996. Biologie der Pilze. – Ulmer, Stuttgart.
Taylor, F. J. R. and Pollingher, U. 1987. Ecology of dinoflagellates: general and marine ecosystems. – Bot. Monogr. 21: 398–502.
Tyler, P. A. 1996. Endemism in freshwater algae with special reference to the Australian region. – Hydrobiologia 336: 1–9.
Vyverman, W. 1996. The Indo-Malaysian North-Australian phycogeographical region revised. – Hydrobiologia 336: 107–120.
Webster, J. 1983. Pilze. Eine Einführung. – Springer.
Wolda, H. 1981. Similarity indices, sample size and diversity. – Oecologia 50: 296–302.
Wolda, H. 1983. Diversity, diversity indices and tropical cockroaches. – Oecologia 58: 290–298.

Appendix A. The ACE (abundance-based coverage estimator) formula

A basic model assumption in the ACE approach and in similarity analysis is that data should be a random (representative) sample. Intuitively, this means that species abundance distributions in data and in community are similar. We tried to meet this assumption by sampling a variety of biotopes and mixing 10 small subsamples to form a composite sample, as discussed in the Section on Material and Methods.

Denote the true number of species richness by S and let f_m be the number of species that occur exactly m times (or in m samples), $m = 1, 2, \ldots$ In the ACE approach, we first select a cut-off point κ, which is 10 by default in the program SPADE. This cut-off is used to separate the observed species into two groups, "rare" and "abundant"; the former group includes those species observed at most 10 times and the latter includes those observed at least 11 times. Only the statistics ($f_1, f_2, \ldots, f_{10}$) are used to estimate the number of missing species, because those abundant species would be observed in almost every sample and thus carry negligible information about the missing species. Define (for the cut-off $\kappa = 10$)

D_{rare}: the number of distinct species for "rare" group; $D_{rare} = \Sigma_{i=1}^{\kappa} f_i$

D_{abun}: the number of distinct species for "abundant" group; $D_{abun} = \Sigma_{i>\kappa} f_i$

$\hat{C}_{rare}$: estimated sample coverage for "rare" group; $\hat{C}_{rare} = 1 - f_1/\Sigma_{i=1}^{\kappa} i f_i$

$\hat{\gamma}_{rare}$: estimated CV (coefficient of variation), which measures the heterogeneity of species abundances;

$\tilde{\gamma}_{rare}$: estimated CV for a highly heterogeneous case.

The ACE for estimating species richness in a moderately heterogeneous community (Chao and Lee 1992) takes the following form:

$$\hat{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2_{rare}$$

$$\text{where } \hat{\gamma^2_{rare}} = \max\left\{ \frac{D_{rare}}{\hat{C}_{rare}} \frac{\Sigma_{i=1}^{\kappa} i(i-1)f_i}{(\Sigma_{i=1}^{\kappa} i f_i)(\Sigma_{i=1}^{\kappa} i f_i - 1)} - 1,\ 0 \right\}$$

and $\max\{a, 0\} = a$ if a is non-negative and $= 0$ if a is negative.

The ACE for a highly heterogeneous case (Chao et al. 1993, Chao 2005), which is called ACE_1 in SPADE, is

$$\tilde{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \tilde{\gamma}^2_{rare}$$

where

$$\tilde{\gamma}^2_{rare} = \max\left\{ \hat{\gamma}^2_{rare}\left(1 + \frac{(1-\hat{C}_{rare})\Sigma_{i=1}^{\kappa} i(i-1)f_i}{\hat{C}_{rare}(\Sigma_{i=1}^{\kappa} i f_i - 1)}\right),\ 0 \right\}.$$

Appendix B. Checking the reliability of data analysis by prediction

Suppose the presence/absence data were collected from n samples and we want to predict the number of unseen species that will be observed in a m additional samples. Let $t = m/n$ and define S(t): the expected number of unseen species that will be observed in a fraction t of the original samples. Chao and Shen (2004) proposed the following predictor

$$\hat{S}(t) = \hat{f}_0[1 - \exp(-t \leftrightarrows f_1/\hat{f}_0)]$$

where $\hat{f}_0 = \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2_{rare} - D_{rare}$, an estimator for unseen species, is obtained from the ACE formula provided in Appendix A. This predictor is featured in Program SPADE (Chao and Shen 2003-2005). For the undescribed data in Namibia, we have $f_1 = 118$ (the number of singlets) and the ACE yields $\hat{f}_0 = 360$; thus the predicted number of unseen undescribed species in an additional set of 5 samples becomes $\hat{S}(t) = 360\ [1 - \exp(-\frac{5 \times 118}{73 \times 360}] = 8.$