ELSEVIER

FEMS Microbiology Ecology 28 (1999) 325–334

FEMS MICROBIOLOGY ECOLOGY

Towards complete biodiversity assessment: an evaluation of the subterranean bacterial communities in the Oklo region of the sole surviving natural nuclear reactor

R.H. Crozier [a,*], P.-M. Agapow [1,a], K. Pedersen [b]

[a] *Centre for Conservation Genetics and Department of Biochemistry and Genetics, La Trobe University, Bundoora, Vic. 3083, Australia*
[b] *Microbiology Section, Department of Cell and Molecular Biology, University of Göteborg, Box 462, SE-405 30 Göteborg, Sweden*

Received 10 July 1998; received in revised form 6 November 1998; accepted 10 November 1998

Abstract

Groundwater bacterial rRNA sequences extracted from the natural nuclear reactor region of Gabon are used to demonstrate the application of phylogenetic methods to biodiversity assessment. Clones were provisionally placed in 'genera' using either the genus of the closest named EMBL entry, or by grouping clones at least 97.5% identical. The community is small, with 24 putative genera under the 'closest-match' criterion and an estimated number of 30.9 (25.8–49.7); estimated genus sample coverage is therefore 78% (48.3–92.8%). There were 36 genera under the 'threshold' criterion, with an estimated number of 87.2 (52.6–193.8), and sample coverage 41.3% (18.6–68.4%) Molecular biodiversity was estimated for all site combinations using genetic diversity (GD: probability of at least two alleles being present in the sequences preserved), and confidence limits derived by standard phylogenetic bootstrap sampling from the sequence dataset. Some combinations with fewer sites preserved GD as well as combinations with larger numbers, although GD is maximised by preserving as many sites as possible depending on choice of site. Some site combinations did not differ significantly in GD preserved, and the conservation value of a site depends on the others selected. The strongest predictor of molecular biodiversity is the observed number of 'closest-match' genera, supporting the higher taxon richness concept for biodiversity. The similarities between sites, and hence the molecular biodiversity characteristics of combinations of them, was associated with similarity in physical characteristics, the availability of organic carbon, and depth below the surface.

Keywords: Conservation genetics; Subterranean bacteria; Molecular biodiversity; Ribosomal DNA sequence; Molecular phylogeny; Oklo natural nuclear reactor

1. Introduction

The Earth has between 3 million and 30 million species [1], presenting a vast array of characteristics. The problem of biodiversity preservation is to maintain the broadest array of these possible. An important aspect of this problem is the assessment of the

* Corresponding author. Fax: +61 (3) 9479-2480; E-mail: r.crozier@gen.latrobe.edu.au

[1] Present address: Department of Biology, Imperial College at Silwood Park, Silwood Park, Berks, UK.

0168-6496/99/$20.00
PII: S0168-6496(98)00116-0

biodiversity present in species groups and, especially, in possible reserve systems. Such assessments should ultimately take into account the long-term results expected of biodiversity preservation policies rather than the current biodiversity levels for the various possible reserves [2].

Species differ in their degree of distinctiveness, and under various philosophical frameworks for the preservation of biodiversity it is desirable to give highest weight to measures that preserve the most diverse set of species [3]. Phylogenetic methods have been proposed as the best way to achieve biodiversity assessment for maximum biodiversity (reviewed in [3]), and Wilson's [4] characterisation of biodiversity as genetic information content leads naturally to the use of molecular phylogenies for this purpose because of their utility for phylogenetic studies and the natural and universal divergence scale they present.

A major problem in biodiversity assessment is the difficulty in collecting sufficient data for reliability, compounded by the current general practice of omitting statistical confidence limits. Surveys are inevitably partial, with considerable effort being deployed to overcome this deficiency with the determination of surrogate measures. The use of indicator groups whose variation in abundance might reflect that of unstudied groups has occasionally, but not always, been successful [3]). Higher taxon richness, in which the numbers of higher categories (e.g. families) rather than species is used as an indicator of species richness shows promise [5]. Furthermore, because higher categories reflect the judgement of systematists in the degree of distinctiveness of the species concerned, higher taxon richness might be a reasonable surrogate for biodiversity in terms of diversity [3]), although there is some evidence that the evolutionary distinctiveness of higher taxa may vary systematically from region to region in some cases [6].

Two chief measures have been proposed for the phylogenetic estimation of biodiversity. 'Phylogenetic diversity' (PD) [7,8] is the length of phylogenetic tree preserved, whereas 'genetic diversity' (GD) [3,9–11] is the probability of preserving at least two alleles in the set of taxa retained. Given that the goal of biodiversity preservation is maintenance of diversity, phylogenetic trees with branch lengths expressed in differences rather than substitutions are more appropriate, and for distant comparisons (such as between Eukarya and Bacteria) differences in gene number should also be considered [3]. Because the values of GD and PD are dominated by the length of the branch connecting a group to the rest of the phylogeny, the results are fairly resistant to variation in the trees returned, including in topology [10].

Phylogenetic methods have so far been applied to groups of perceived high conservation significance, cranes being a noteworthy example [2,12,13]. The application of phylogenetic methods to the assessment of total biodiversity assessment has encountered the same difficulty, in greater degree, of all other data types – the task of collection. However, bacteria represent a prime target for molecular assessment of biodiversity, not only because of possible greater ease of data collection than for larger organisms, but also because biodiversity information on bacteria is essentially unobtainable by any other means because the great majority of them cannot be cultured. The size of the bacterial biota may be substantial, with 4000 species estimated per gram of Norwegian forest soil through molecular methods [14]. Given also the likely utilitarian benefits of this poorly known major part of the biota [15], bacterial biodiversity should become a major field of conservation effort.

Although there are diverse current efforts to discover new bacteria from groups which cannot readily be cultured (e.g. [16]), and to estimate the diversity of bacterial communities in specific sites [17], datasets comparing several sites appear to be rare. One such dataset comprises 126 16S rRNA sequences [18] from bacteria from groundwaters sampled through five boreholes in the vicinity of the world's last known natural nuclear reactor at Bangombé, in the Oklo region of Gabon [19]. Although initiated in connection with the need to develop nuclear power as a relatively environmentally benign energy source [20], these data also provide a prototype for the analysis of biodiversity preservation through maintenance of the regions sampled, and for the estimation of the total community size.

2. Materials and methods

The Oklo region of Gabon has a stratum presenting several nuclear reactors due to the ancient con-

centration of uranium ore through sedimentary processes [21]. All but one of these have been destroyed by mining operations, the survivor being at Bangombé [19]. These reactors went critical approximately 2 billion years ago and today are represented only by fossilised remnants.

Samples were collected from screened sections of five boreholes (BAX01, BAX02, BAX03, BAX04 and BAX07), and sequences obtained as described previously [18]. Briefly, the boreholes were at various depths along a 162 m SW–NE transect. BAX01 was the deepest (96–105 m, in sandstone) and BAX07 the shallowest (4.5–6.5 m, in the overburden). BAX02 sampled the 27.2–33.9 m range. BAX03 and BAX04 were 5 m apart and sampled the reactor itself with surrounding strata (11.9–12.5 m), and layers just above the reactor (8.9–10.2 m), respectively. The boreholes were chosen to represent different environments in and around the reactor, each having different groundwater and geological characteristics [18]. A cross-sectional view of the site has been presented previously [18].

Universal PCR-primers were chosen to detect bacterial, archaeal and eukaroyote sequences. These methods yield few, if any, contamination or artefactual problems [22,23].

Sequences were assigned tentatively to genera using two methods: according to the genus of the named EMBL database sequence they were closest to, and by grouping together clones whose sequences were at least 97.5% identical. We term these 'closest' and 'threshold' genera. The percent identity of a Bangombé clone to its closest database match ranged from 77.4 to 99.4%. There is no accepted value of the percent identity at which two 16S rRNA sequences can be concluded to belong to the same genus or species, but rather the identity value can differ markedly for different genera [24] and also according to whether total or partial 16S rRNA genes are compared. It has been suggested, based on comparisons of rDNA sequences and on DNA–DNA reassociation, that sequences do not come from the same species if they show less than 97.5% identity of 16S rDNA sequence. At higher identity values, species identity must be confirmed using DNA–DNA hybridisation [25]. We chose 97.5% as a conservative measure enabling us to compare two methods of grouping sequences.

Coverage-based methods (reviewed by [26,27]) were used to estimate the numbers of different clones and genera. Briefly, the proportion of single occurrences in samples is used to estimate the number of species not occurring in the sample. The coefficient of variation of the observed taxa is used to modify the initial coverage-based estimate [28]. Simulation indicates that although the resulting estimates tend to be too low, they are generally close to the actual number of species [29]. Although a martingale estimation procedure is available [29], it requires knowledge of the order in which taxa are encountered and we did not feel confident of recovering this information from laboratory records. The coefficient of variation, and hence the size of the confidence interval, is reduced by temporarily removing very common taxa from the calculations and restoring them later [29].

We used the C program CHAO kindly supplied by Dr Anne Chao for estimating the numbers of taxa and the associated standard errors. These values were then used to generate asymmetrical confidence limits [30].

The 126 sequences were aligned as described previously [18]. 1000 bootstrap subsamples of the dataset were made using the program SEQBOOT in the phylogeny inference package PHYLIP [31] implemented on a Power Macintosh computer, and we used these to derive 1000 Jukes–Cantor distance matrices [32] using the PHYLIP program DNADIST. The distance matrices were used to derive 1000 neighbour-joining phylogenetic trees [33] using the PHYLIP program NEIGHBOR, and the resulting trees were kept in condensed format in a treefile produced by the program. Visualisation of trees used the program TREEVIEW FOR POWERMACINTOSH (Roderick Page, 1997).

The treefile with 1000 trees was read by the Macintosh C program CONSERVE 3.1.2 (available from R.H.C.'s section of the worldwide web site http://www.gen.latrobe.edu.au). CONSERVE allows the estimation of both GD and PD, the conversion of Jukes–Cantor distances to *p*-distances (distances involving sequence differences and not estimates of substitution rates), and the setting of 95% confidence limits using standard bootstrap methods [34,35]. We also used the capability of the program to express all biodiversity values in terms of the per-

Table 1
Characteristics of all combinations of the five Bangombé sites

No. sites preserved	Sample size	Clones observed	Estimated clones	'Closest' genera observed	Estimated 'closest' genera	'Threshold' genera observed	Estimated 'threshold' genera	Genetic diversity
5	126	40	64.68 < 111.38 < 246.46	24	25.85 < 30.89 < 49.69	36	52.61 < 87.20 < 193.84	100
4	116	37	59.63 < 103.68 < 233.51	22	23.51 < 27.95 < 45.51	33	46.95 < 77.49 < 174.85	98.73 < 98.74 < 98.75
4	99	34	54.01 < 96.24 < 227.61	22	24.52 < 31.09 < 54.8	31	43.42 < 71.80 < 165.07	97.93 < 97.95 < 97.97
4	97	34	49.25 < 80.89 < 178.15	21	22.12 < 25.64 < 40.28	30	38.63 < 59.10 < 128.18	96.50 < 96.54 < 96.57
4	96	35	52.55 < 88.08 < 195.55	23	25.04 < 30.39 < 49.74	33	46.96 < 76.21 < 166.70	98.540 < 98.6 < 98.57
4	96	29	36.93 < 56.35 < 123.32	20	20.62 < 23.08 < 35.27	27	33.28 < 49.82 < 109.94	98.23 < 98.24 < 98.26
3	89	31	53.69 < 104.87 < 271.43	20	22.38 < 28.82 < 52.75	28	38.97 < 65.35 < 155.19	96.04 < 96.07 < 96.11
3	87	31	47.54 < 83.64 < 198.52	19	20.66 < 25.65 < 45.67	27	35.50 < 57.27 < 134.82	94.16 < 94.22 < 94.27
3	70	28	39.77 < 67.01 < 157.34	18	19.77 < 25.00 < 45.62	25	31.80 < 49.34 < 112.08	92.73 < 92.80 < 92.86
3	86	32	52.69 < 89.38 < 191.12	21	23.14 < 28.93 < 50.44	30	43.05 < 72.79 < 170.27	96.84 < 96.87 < 96.90
3	69	26	38.91 < 73.12 < 198.03	18	23.41 < 45.21 < 129.04	24	33.71 < 60.18 < 158.81	92.92 < 92.98 < 93.04
3	67	28	40.68 < 70.83 < 172.71	19	20.85 < 26.03 < 45.76	26	31.99 < 43.83 < 79.06	92.06 < 92.13 < 92.19
3	86	26	34.09 < 55.05 < 130.35	18	18.76 < 21.65 < 35.59	24	30.26 < 47.96 < 115.74	96.35 < 96.38 < 96.41
3	69	23	28.74 < 44.75 < 105.38	18	18.97 < 22.38 < 37.87	22	25.48 < 33.73 < 61.49	95.30 < 95.34 < 95.38
3	67	23	25.67 < 32.34 < 55.72	17	17.17 < 18.18 < 25.33	21	22.99 < 28.49 < 49.18	91.55 < 91.62 < 91.68
3	66	24	29.23 < 40.28 < 74.72	19	20.00 < 23.31 < 37.66	24	29.23 < 40.28 < 74.72	96.17 < 96.20 < 96.24
2	60	25	41.49 < 83.59 < 233.22	15	16.61 < 21.79 < 43.69	22	29.33 < 49.68 < 126.57	89.24 < 89.32 < 89.41
2	59	23	44.61 < 111.34 < 384.17	16	24.52 < 58.09 < 223.86	21	31.38 < 64.57 < 203.80	89.54 < 89.62 < 89.69
2	57	24	35.05 < 58.65 < 132.65	16	17.51 < 22.29 < 42.16	22	25.83 < 34.6 < 63.41	87.22 < 87.32 < 87.41
2	40	19	32.06 < 72.19 < 235.71	12	17.34 < 42.94 < 191.28	17	26.02 < 56.28 < 188.00	78.05 < 78.19 < 78.32
2	59	19	23.59 < 38.24 < 99.56	15	15.85 < 19.12 < 34.93	18	20.59 < 27.68 < 54.11	92.10 < 92.16 < 92.22
2	57	20	24.15 < 37.78 < 96.15	15	15.71 < 18.61 < 33.34	18	20.55 < 27.68 < 54.69	87.54 < 87.63 < 87.72
2	40	16	17.97 < 23.78 < 46.70	14	14.53 < 16.78 < 28.64	15	16.31 < 20.65 < 39.41	84.96 < 85.07 < 85.17
2	56	21	30.00 < 56.47 < 160.84	17	18.73 < 23.98 < 45.17	21	30.00 < 56.47 < 160.84	93.66 < 93.71 < 93.76
2	39	14	15.73 < 21.62 < 47.64	13	14.12 < 18.45 < 39.50	14	15.73 < 21.62 < 47.64	87.13 < 87.22 < 87.31
2	37	17	19.16 < 25.44 < 49.95	15	15.67 < 18.34 < 31.66	17	19.16 < 25.44 < 49.95	84.28 < 84.38 < 84.49
1	30	14	24.62 < 65.37 < 262.55	7	8.06 < 15.49 < 74.85	12	15.91 < 31.98 < 114.22	55.77 < 55.96 < 56.15
1	30	12	14.71 < 26.28 < 87.13	10	10.41 < 12.46 < 24.75	11	12.26 < 17.03 < 39.81	77.95 < 78.08 < 78.21
1	29	10	17.16 < 29.29 < 61.79	10	11.99 < 20.29 < 63.10	10	11.99 < 20.29 < 63.10	80.52 < 80.63 < 80.74
1	27	13	14.72 < 20.39 < 44.74	12	13.07 < 17.09 < 36.24	13	14.72 < 20.39 < 44.74	75.60 < 75.73 < 75.86
1	10	6	6.92 < 12.22 < 48.27	6	6.92 < 12.22 < 48.27	6	6.92 < 12.22 < 58.27	57.39 < 57.56 < 57.72

The observed and estimated numbers of clones and genera are given, together with the 95% confidence intervals of the latter. Genera are determined on the basis of EMBL database entry most closely matching the sequence ('closest' criterion) or clones were regarded as congeneric if they were at least 97.5% identical. Genetic diversity and its 95% confidence intervals are shown, given as percentage of the diversity preserved for all five sites.

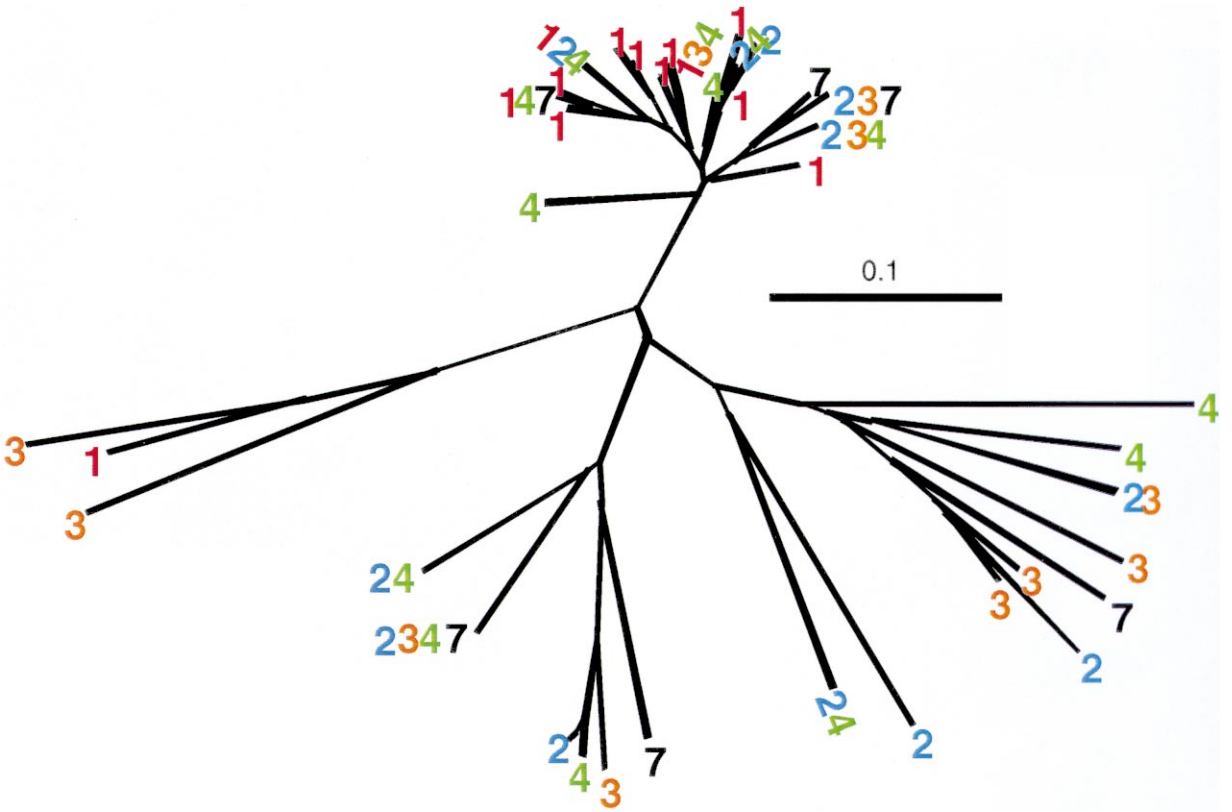

Fig. 1. Neighbour-joining phylogenetic tree of the observed 16S rRNA sequence derived using a matrix of Jukes–Cantor distances, with the branch lengths converted to *p*-distances after the tree was inferred. The boreholes from which each clone was derived are shown; some clones occurred in more than one borehole. The scale indicates sequence difference per base.

centages of the biodiversity of the complete set of sites.

The program allows menu-driven addition and deletion from the analysis of all species as a group at a site rather than through piecemeal individual treatment (although this is allowed), and we used this capability in this analysis. The program also allows an indefinitely large dataset depending on the capacity of the computer used.

For multiple ([35], pp. 610–634) and stepwise multiple ([35], pp. 654–664) regression analysis we used StatView 4.5 (Abacus Concepts), using as the dependent variable GD for each combination of sites and as predictor variables the number of sites preserved, the sample size, the number of clones, the observed and estimated numbers of 'closest' genera and the observed and estimated numbers of 'threshold' genera. For curve-fitting and graphing we used CricketGraph 1.5.3 (Computer Associates).

3. Results

The observed and estimated numbers of clones and bacterial genera at the various possible combinations of sites, together with molecular biodiversity estimates (GD), and the confidence limits for these quantities, are given in Table 1. To save space, we present only GD, the measure explicitly designed for molecular diversity, and not PD, because if desired the latter can be derived for these data by using the empirically determined equation $PD = 3.881 \times 10^{0.014GD}$.

The phylogenetic tree of the sequences used, with

distances converted to *p*-distances and showing the bore-hole origins, is given in Fig. 1.

The relationship between numbers of sites preserved and GD is given in Fig. 2. For each category of number of sites preserved the combination preserving the most biodiversity is indicated. Given choice of sites within each category of number of sites preserved, preserving more sites is always best. However, the GD values overlap considerably between categories, with for example the combination of boreholes 3 and 4 preserves more GD than the three-site combinations 1+3+7, 1+2+7, 1+4+7 and 2+4+7, and the single borehole 3 preserves more GD than the two-site combination 1+7.

These notes indicate how GD considerations apply to the principle of complementarity. Managers establishing a reserve system would best select not the array of sites which individually retain most species, but that group of sites which retains most diversity. In the present case, the two sites which by themselves maintain most biodiversity are 2 (78.08% of the total) and 3 (80.63%), but of the two-site combinations it is not 2 and 3 (92.16%) but 3 and 4 (93.71%) which maintains the most biodiversity due to the greater complementarity between 3 and 4 as against between

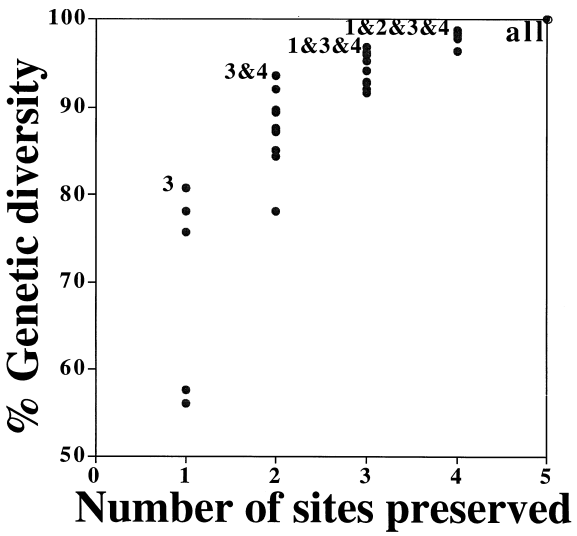

Fig. 2. Genetic diversity (given as a percentage of the value for the complete dataset) plotted as a function of the number of sites preserved for various combinations of sites. The combinations of sites within each class preserving the most biodiversity are shown.

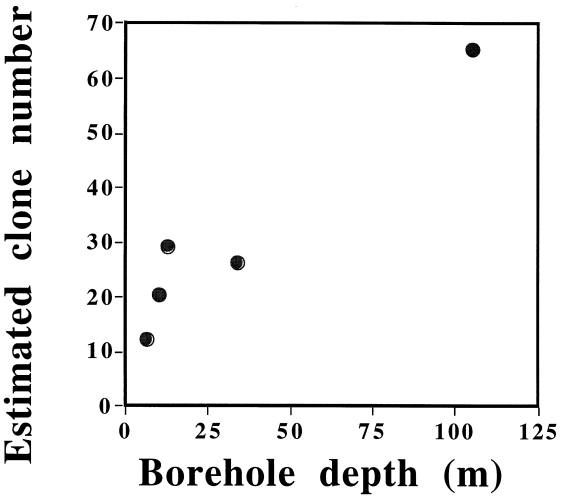

Fig. 3. Estimated numbers of clones for sampling stations at different depths.

2 and 3. Of course, such decisions must involve not only the statistical significance of the difference (3 and 4 would maintain significantly more biodiversity than 2 and 3), but also cost-benefit analyses [2].

In two instances, two combinations could not be statistically separated in terms of GD preserved in that the point estimates of each were within the confidence limits of the other: site 2 versus sites 1+7, and combination 2+3 versus 1+4+7. Borehole BAX03 is in each of the most biodiverse combinations within categories, with BAX04 the next most important contributor to GD.

Although only ten sequences and six clones were sampled from BAX07, compared with 30 sequences and 14 clones from BAX01, the latter is a significantly less genetically diverse single site than BAX07. Perusal of Fig. 1 shows that the BAX07 clones show a wider phylogenetic range than those from BAX01, even though the latter are more numerous. There is a marked increase in estimated clone numbers with depth (Fig. 3), but other quantities show weakly negative relationships.

Multiple regression analysis yielded three statistically significant predictor variables: the observed number of 'closest' genera ($P = 0.0002$), the number of sites preserved ($P = 0.0306$) and the sample size ($P = 0.0329$). All other predictor variables had non-significant coefficients. Listing variables in order of

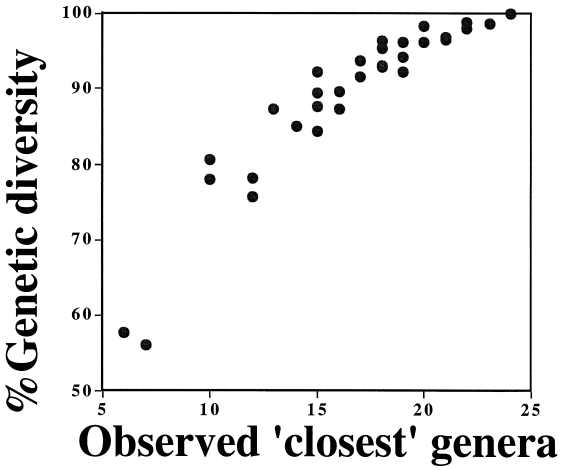

Fig. 4. Genetic diversity (given as a percentage of the value for the complete dataset) plotted as function of the observed number of genera in the various possible combinations of sites preserved; generic assignment was made according to that of the closest match in EMBL. The observed number of genera is the strongest predictor of the amount of biodiversity preserved by combinations of sites.

statistical significance, the multiple regression formula found was GD = 44.528+4.718 (observed number of 'closest' genera)−5.260 (number of sites preserved)+0.252 (sample size)−2.373 (observed number of 'threshold' genera)+0.174 (estimated number of 'threshold' genera)−0.061 (estimated number of clones)+0.689 (observed number of clones)−0.031 (estimated number of threshold genera).

Forward stepwise regression retained only the observed numbers of 'closest' (F-to-remove = 71.102) and 'threshold' genera (F-to-remove 19.860). Backward stepwise regression retained the observed 'closest' (F-to-remove = 93.210) and 'threshold' (F-to-remove = 23.131) genera, the sample size (F-to-remove = 6.555) and the number of sites preserved (F-to-remove = 6.333).

The observed number of 'closest' genera is thus consistently identified as the most important predictor variable for molecular biodiversity, and the relationship between the two quantities is shown in Fig. 4.

The sequences obtained are deposited in EMBL under accession numbers X91173-88, X91191 and X91271-96.

4. Discussion

The GD and PD values can be interpreted in terms of differences with respect to geological characteristics of the studied strata and available organic carbon in the groundwater which was found to correlate significantly with the observed total numbers of bacteria [18]. BAX01 groundwater had the highest content of organic carbon material (6.6 mg l^{-1}), followed by BAX03 (4.1 mg l^{-1}). BAX01 also had the highest concentration of bacteria (5.8×10^5 cells ml^{-1}) followed by BAX03 (5.5×10^5 cells ml^{-1}). Of these two boreholes, BAX03 stands out as the most critical site together with BAX04 with respect to the preservation of GD. It is also clear from the data that the importance of a site varies with the others with which it is grouped. For example, BAX02 does not enter into the topmost combination among three-site groups (BAX01, BAX03, BAX04), but the combination BAX02 and BAX03 is second among the two-site groups and ahead of all those containing BAX01.

Although details about the local flow around the reactor are uncertain, data on pH, conductivity, organic carbon and total number of bacteria suggest that BAX02 and BAX04 are hydraulically connected [18]. It can therefore be expected that they also share a similar GD. Hence, adding BAX02 to a group containing BAX04 should have a lesser influence on GD than adding any of the other boreholes, and indeed this is the case.

BAX03 intersects the nuclear reactor remnant which is no thicker than 18 cm and stretches between a green pelites layer on top and a sandstone one below. The section sampled was 0.6 m thick and it is therefore likely that BAX03 sampled groundwater from three very different geological strata simultaneously: green pelites, the nuclear reactor and sandstone. This is also true for BAX04, which sampled a 1.3-m-long section through the pelites and the overburden. The probability of sampling many different species is reasonably expected to be higher if many different environments are sampled as occurred for both BAX03 and BAX04 as against BAX02 and BAX01 (which both sampled homogeneous strata at greater depths). BAX07 sampled shallow overburden and this site probably accesses a large microbial biodiversity with infiltrating

groundwater carrying many different surface communities. The increase in clone numbers but not GD with depth suggests that compared to sites closer to the surface the deepest site (BAX01) presents a narrower ecological range which permits a narrow phylogenetic range of organisms specialised to live there. Paradoxically more diversification within the lineages present was observed for BAX01, which may indicate that the community may have been adapting relatively undisturbed to its deep and stable environment for a long period of time.

The higher predictive value of 'closest' compared to 'threshold' genera can be understood in terms of the fact that rather few groups emerged according to the threshold criterion: the 40 clones found were grouped into 24 'closest' genera, but into 36 'threshold' genera, indicating that under the more stringent criterion, most clones were not grouped with others. The small sample size for BAX07 is an adequate explanation of the selection of sample size in two of the three analyses as a significant predictor variable.

The higher taxon richness approach is supported by these data, in that the observed number of 'closest' genera emerged as by far the strongest predictor of GD. With respect to application of the concept to eukaryotes, it would seem reasonable to proceed by calibrating selected higher taxa with respect to divergence, to allow quantitative biodiversity assessment to proceed prior to the development of technology allowing rapid sequence generation from diverse sources. The chief objective of this calibration would be to determine the taxonomic level best used in different major groups. The growth of sequence information in genomic databases, which are unmatched by any similar morphological databases [36], is likely to yield broader understanding of the evolution of complexity in organisms, facilitate the identification of unculturable organisms, and assist the use of molecular phylogenies in conservation.

The findings above may be considered in the light of simulation results [37] indicating that preservation of species at random is little worse in the conservation of 'evolutionary history' than using phylogenetic criteria; these results indicated that 95% of the simulated species could be extirpated at random while protecting 80% of the length of the phylogenetic tree. Our results appear to differ from these, in that conservation of only one site preserves in all instances much more than 5% of the estimated clones or genera, but generally much less than 80% of the biodiversity (Table 1). This apparent difference in findings reflects the association of lineages non-randomly with site in the Bangombé data, emphasising the need to consider whole habitats rather than individual constituent species.

The size of the bacterial community in the Bangombé groundwaters is quite small compared with the numbers expected in surface soils [14], with the upper confidence limit of number of 'closest' genera being only 49.69 (Table 1). In other investigations similar to the Bangombé study, larger molecular biodiversities have been observed, i.e. 110 specific clone groups out of 313 sequenced clones from 15 deep granitic boreholes at the Äspö hard rock laboratory [22,38,39]. Both these investigations and the Bangombé study [20] have not exhausted the sequences to be found because new sequences were found in nearly every additional sample. The data collection effort will therefore clearly have to be scaled up significantly for the study of most communities, requiring the application of automated procedures. Bacteria are likely to be the first group of organisms for which such automated biodiversity assessment is practicable (although other soil microbiota might also be so surveyed); while as a major constituent of the community they deserve assessment in their own right in addition to their value as indicators.

The analyses above indicate the coming ability for complete biodiversity assessment, but for microbial communities generally it is often assumed that many species occur in most habitats in low numbers, needing only appropriate conditions to multiply to detectable levels [40–42]. Under this view, molecular approaches such as we have described would be useful for: (a) uncovering novel taxa among unculturable forms [17,43–47]; or (b) surveying environmental conditions, reflected by which microorganisms are actively multiplying and hence most readily detectable by PCR-based methods [48–52]. However, it has been pointed out that, because of the past reliance on culturable organisms, the numbers of species of microorganisms may be several orders of magnitude higher than commonly assumed [46,47,53], leading to the conclusion that whether or not there are endangered bacterial species in the same way as there are

endangered metazoa and metaphytes remains an open question until more data accrue [53].

A caveat to be mentioned is that, although the confidence limits to GD we obtained are gratifyingly small, these pertain to the particular set of sequences studied. It is not yet clear how to take account of sequences not obtained. For example, bootstrap samples will, because the analysis rests on differences between sequences, always be 'less diverse' than the complete sample if they differ from it. Further analytical or simulation work on this problem is desirable.

Acknowledgments

We thank Niels Becker, John Bunge and especially Anne Chao for statistical advice on estimating clone and genera numbers, Anne Chao for sending her estimation program CHAO, and for support of this work the Australian Research Council (R.H.C.) and the Swedish Nuclear Fuel and Waste Management and the Swedish Natural Science Research Council (K.P.).

References

[1] May, R.M. and Nee, S. (1995) The species alias problem. Nature 378, 447–448.
[2] Weitzman, M.L. (1993) What to preserve? An application of diversity theory to crane preservation. Q. J. Econom. 108, 157–183.
[3] Crozier, R.H. (1997) Preserving the information content of species: genetic diversity, phylogeny and conservation worth. Annu. Rev. Ecol. Syst. 28, 243–268.
[4] Wilson, E.O. (1992) The Diversity of Life. Harvard Univ. Press, Cambridge, MA.
[5] Linder, H.P. and Midgley, J.J. (1994) Taxonomy, compositional biodiversity and functional biodiversity of fynbos. S. Afr. J. Sci. 90, 329–333.
[6] Gaston, K.J. and Blackburn, T.M. (1996) The tropics as a museum of biological diversity: analysis of the New World avifauna. Proc. R. Soc. Lond. B. Biol. Sci. 263, 63–68.
[7] Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10.
[8] Faith, D.P. (1994) Phylogenetic diversity: a general framework for the prediction of feature diversity. In: Systematics and Conservation Evaluation (Forey, P.L., Humphries, C.J. and Vane-Wright, R.I., Eds.), pp. 251–268. Clarendon Press, Oxford, UK.
[9] Crozier, R.H. (1992) Genetic diversity and the agony of choice. Biol. Conserv. 61, 11–15.
[10] Crozier, R.H. and Kusmierski, R.M. (1994) Genetic distances and the setting of conservation priorities. In: Conservation Genetics (Loeschcke, V., Tomiuk, J. and Jain, S.K., Eds.), pp. 227–237. Birkhäuser Verlag, Basle.
[11] Crozier, R.H. (1997) A genetic diversity approach to conservation: genetic similarities and differences between species. Proc. Assoc. Adv. Anim. Breed. Genet. 12, 624–632.
[12] Weitzman, M.L. (1992) On diversity. Q. J. Economics, 107 363–405.
[13] Krajewski, C. (1994) Phylogenetic measures of biodiversity: a comparison and critique. Biol. Conserv. 69, 33–39.
[14] Torsvik, V., Goksøyr J. and Daae, F.L. (1990) High density in DNA of soil bacteria. Appl. Environ. Microbiol. 56, 782–787.
[15] Istock, C.A., Bell, J.A., Ferguson, N. and Istock, N.L. (1996) Bacterial species and evolution: theoretical and practical perspectives. J. Ind. Microbiol. 17, 137–150.
[16] Ohkuma, M., Noda, S., Horikoshi, K. and Kudo, T. (1995) Phylogeny of symbiotic methanogens in the gut of the termite *Reticulitermes speratus*. FEMS Microbiol. Lett. 134, 45–50.
[17] Bintrim, S.B., Donohue, T.J., Handelsman, J., Roberts, G.P. and Goodman, R.M. (1997) Molecular phylogeny of Archaea from soil. Proc. Natl. Acad. Sci. USA 94, 277–282.
[18] Pedersen, K., Arlinger, J., Hallbeck, L. and Pettersson, C. (1996) Diversity and distribution of subterranean bacteria in groundwater at Oklo in Gabon, Africa, as determined by 16S rRNA gene sequencing. Mol. Ecol. 5, 427–436.
[19] Gauthier-Lafaye, F., Blanc, P.L., Bruno, J., Griffault, L., Ledoux, E., Louvat, D., Michaud, V., Montoto, M., Oversby, V., Delvillar, L.P. and Smellie, J. (1997) The last natural nuclear fission reactor. Nature 387, 337.
[20] Pedersen, K. (1996) Investigations of subterranean bacteria in deep crystalline bedrock and their importance for the disposal of nuclear waste. Can. J. Microbiol. 42, 382–391.
[21] Cowan, G.A. (1976) A natural fisson reactor. Sci. Am. 235, 36–47.
[22] Pedersen, K., Hallbeck, L., Arlinger, J., Erlandson, A.-C. and Jahromi, N. (1997) Investigation of the potential for microbial contamination of deep granitic aquifers during drilling using 16S rRNA gene sequencing and culturing method. J. Microbiol. Meth. 30, 179–192.
[23] Huber, R., Burggraf, S., Mayer, T., Barns, S.M., Rossnagel, P. and Stetter, K.O. (1995) Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. Nature 376, 57–58.
[24] Fox, G.E., Wisotzkey, J.D. and Jurtshuk, P. (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. J. Syst. Bacteriol. 42, 166–170.
[25] Stackebrandt, E. and Goebel, B.M. (1994) Taxonomic note: a place for DNA–DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846–849.
[26] Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. J. Am. Stat. Assoc. 88, 364–373.
[27] Bunge, J., Fitzpatrick, M. and Handley, J. (1995) Comparison

of three estimators of the number of species. J. Appl. Stat. 22, 45–59.
[28] Chao, A. and Lee, S.-M. (1992) Estimating the number of classes via sample coverage. J. Am. Stat. Assoc. 87, 210–217.
[29] Chao, A., Yip, P. and Lin, H.-S. (1996) Estimating the number of species via a martingale estimating function. Stat. Sin. 6, 403–418.
[30] Chao, A. (1989) Estimating population size for sparse data in capture–recapture experiments. Biometrics 45, 427–438.
[31] Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, Univ. of Washington, Seattle.
[32] Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In: Mammalian Protein Metabolism. (Munro, N.H. Munro, Ed.), pp. 21–123. Academic Press, New York.
[33] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenies. Mol. Biol. Evol. 4, 406–425.
[34] Efron, B. and Tibshirani, R. (1993) An Introduction to the Boostrap. Chapman and Hall, New York.
[35] Sokal, R.R. and Rohlf, F.J. (1995) Biometry. The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York, NY.
[36] Leipe, D.D. (1996) Biodiversity, genomes, and DNA sequence databases. Curr. Opin. Genet. Dev. 6, 686–691.
[37] Nee, S. and May, R.M. (1997) Extinction and the loss of evolutionary history. Science 278, 692–694.
[38] Pedersen, K., Arlinger, J., Ekendahl, S. and Hallbeck, L. (1996) 16s rRNA gene diversity of attached and unattached bacteria in boreholes along the access tunnel to the Äspö hard rock laboratory, Sweden. FEMS Microbiol. Ecol. 19, 249–262.
[39] Pedersen, K. (1997) Microbial life in deep granitic rock. FEMS Microbiol. Rev. 20, 399–414.
[40] Finlay, B.J., Corliss, J.O., Esteban, G. and Fenchel, T. (1996) Biodiversity at the microbial level: the number of free-living ciliates in the biosphere. Q. Rev. Biol. 71, 221–237.
[41] Fenchel, T., Esteban, G.F. and Finlay, B.J. (1997) Local versus global diversity of microorganisms: cryptic diversity of ciliated protozoa. Oikos 80, 220–225.
[42] Finlay, B.J., Maberly, S.C. and Cooper, J.I. (1997) Microbial diversity and ecosystem function. Oikos 80, 209–213.
[43] Pace, N.R. (1996) New perspective on the natural microbial world: molecular microbial ecology. ASM News 62, 463–470.
[44] Huang, W.M. (1996) Bacterial diversity based on type II DNA topoisomerase genes. Annu. Rev. Genet. 30, 79–107.
[45] Bowman, J.P., McCammon, S.A., Brown, M.V., Nichols, D.S. and McMeekin, T.A. (1997) Diversity and association of psychrophilic bacteria in Antarctic sea ice. Appl. Env. Microbiol. 63, 3068–3078.
[46] Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. Science 276, 734–740.
[47] DeLong, E.F. (1997) Marine microbial diversity: the tip of the iceberg. Trends Biotechnol. 15, 203–207.
[48] Foissner, W. (1997) Protozoa as bioindicators in agroecosystems, with emphasis on farming practices, biocides, and biodiversity. Agric. Ecosyst. Environ. 62, 93–103.
[49] Smit, E., Leeflang, P. and Wernars, K. (1997) Detection of shifts in microbial community structure and diversity in soil caused by copper contamination using amplified ribosomal DNA restriction analysis. FEMS Microbiol. Ecol. 23, 249–261.
[50] Bowman, J.P., Brown, M.V. and Nichols, D.S. (1997) Biodiversity and ecophysiology of bacteria associated with Antarctic sea ice. Antarctic Sci. 9, 134–142.
[51] Acinas, S.G., Rodríguez-Valera, F. and Pedrós-Alio, C. (1997) Spatial and temporal variation in marine bacterioplankton diversity as shown by RFLP fingerprinting of PCR amplified 16s rDNA. FEMS Microbiol. Ecol. 24, 27–40.
[52] Giovannoni, S.J., Rappé, M.S., Vergin, K.L. and Adair, N.L. (1996) 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. Proc. Natl. Acad. Sci. USA 93, 7979–7984.
[53] Staley, J.T. (1997) Biodiversity: are microbial species threatened? Commentary. Curr. Opin. Biotechnol. 8, 340–345.