MICROBIAL TYPING FOR MANAGEMENT OF REMEDIATION IN CONTAMINATED SOILS

Almeida J.S.[1,2], Barreto Crespo, M.T.[1], Figueiredo Marques, J.J.[1,3], Noble P.A.[4], MacNaughton, S.J.[5], Stephen, J.R.[5], White, D.C.[5,6], Carrondo, M.J.T.[1,2]

1) Instituto de Tecnologia Química e Biológica / UNL and Instituto de Biologia Experimental e Tecnológica, Apartado 127, 2780 Oeiras, Portugal
2) Faculdade de Ciências e Tecnologia / UNL, 2825, Monte da Caparica, Portugal
3) Estação Agronómica Nacional / INIA, Quinta do Marquês, 2780 Oeiras, Portugal
4) Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC 29208, USA
5) Ctr for Environmental Biotechnology / Univ. Tennessee, Knoxville, TN 37932-2575, USA
6) Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6036, USA

Abstract

The development of effective methods to monitor soil pollution levels is of critical importance when deciding on the usage of contaminated lands. In addition, it is crucial to anticipate the effect of accelerated microbial activity (by nutrient addition and/or inoculation) and the extent of natural degradation to evaluate alternative management scenarios. It has been shown before, by these and other authors, that both goals may be simultaneously attained by using physiological, biochemical, or genetic typing of the whole microbial community. The application of microbial community typing methods (MCT) requires the development of a centralised database relating contamination levels to microbial community profiles. Once a critical amount of information is accumulated the microbial typing method can be associated with the level of contamination and the rate of remediation. Data digging methodologies based on neural networks (NN), an artificial intelligence technique, were successfully used to deconvolute the association. MCT methods, namely carbon source usage, signature lipid biomarker (PLFA), and PCR amplification followed DGGE are compared and their interpretation by NN is described. Some MCT are very cost effective and only require basic training to be implemented. Other MCT are amenable to automation and can be applied to samples stored and shipped to a central location.

Introduction

Monitoring microbial mediated processes in complex matrices, such as bioremediation in soil, faces two particular challenges. First, natural Microbial communities are very partially accessible as less than 1% of soil bacteria detected by direct count are culturable (Skinner *et al.* 1952, Bakken 1985). Second, nutrient cycles are implemented by fluid associations of different organisms defining an entangled web of interdependencies. As a consequence, process dynamics in ecosystems are hard to predict and manipulate on the basis of mechanistic models (Pahl-Wostl 1995). An additional element of complexity is the persistence of spatial heterogeneity, even when the physical characteristics of the site suggest a homogeneous soil compartment (Di Gregorio 1997, Hastings 1994).

specific component of the system being monitored. The target component may be a class of biochemical compounds, the genetic identity of dominant populations, or the functional capabilities of the overall community. Since that partial information is going to be used to portrait the overall system, it is crucial that it may be as representative as possible.

The analysis of phospholipid fatty acids (PLFA) is a balanced biochemical profiling technique that accesses the overall microbial community composition (White 1996). The technique calls for extraction of lipophilic components (Macnaughton 1997) followed by Gas Chromatography with the subsequent identification of individual PLFA by having Mass Spectrometry in line. It has been demonstrated that the pattern of PLFAs contains information about community microbial identity, functional capabilities and physiological status (Almeida 95). This conclusion is also justified by the fact that PLFAs are the major component of cellular membranes. Although, several other MCT targeting biochemical composition have been applied to soil samples (e.g. Quinones, diglycerides and PHA content) PLFA are arguably the most popular choice, with a resolution as high as ~ 10 femtomoles (White 1996).

The method of choice to profile microbial community identity consists of PCR amplification of rDNA fragments followed by denaturating gel gradient electrophoresis (DGGE; Kowalchuck 1997, Stephen 1998). The isolated DGGE bands can be sequenced and compared to a database of previously characterised sequences. This technique allows the phylogenetic identification of the dominant species present. Alternative techniques include probing for selected genes or RNA sequences (White 1997, Kawaharasaki 1998) and the analysis of restriction fragment length polymorphism (RFLP; Liu 1998). Recent advances in the development of gene chips are augmenting the potentialities of genetic profiling techniques to include functional analysis, which is achieved by targeting mRNA fragments recovered from soil samples (Voordouw 1998).

A large variety of methodologies for functional profiling for MCT exist that have been developed as an extension of the classical microbiology techniques for classification of pure culture isolates (Logan 1994). The use of substrate utilisation patterns in particular became very widespread, in part due to the convenience of commercial substrate utilisation galleries (e.g. BIOLOG^{TM} , API-Biomerieux) . Although its usefulness as a research tool has been questioned (Smalla 1998), if properly used they provide a reproductive and cost effective fingerprint technique easily applied to soil and compost samples (Garland 1996).

Finally, a large amount of information available by simple observation of the field site is often overlooked and fails to be recorded in databases. Although not part of a MCT proper, the field worker is often aware of valuable correlation between environmental and process parameters ranging from the presence of indicator species to the soil/compost material appearance (texture/colour/odour). The subjective nature of this information has precluded its inclusion in statistical correllations. However, artificial learning techniques are not hampered by the same limitations and can use implicit information to produce reliable predictions as will be discussed below.

Test selection

The selection of carbon sources to use, DGGE bands to sequence, genes to probe or the required resolution for the PLFA profiling, has to be made such that sample discrimination is maximised.

defines information of a set of measures as the number of different states resolved. The determination of information content is appropriate for profiles of binary tests such as galleries of substrate utilisation. However, for a broader utilisation of the concept, entropy (S) will be used instead (eq.1); N is the number of observed test outcomes, and pi is the corresponding frequency.

$$S = -\sum_{i=1}^{N} pi \cdot \log_e(pi) \qquad (1)$$

Parameter selection is an iterative process that proceeds until the required discrimination between samples is reached. The optimal combination for a given number of tests or parameters does not necessarily include the optimal subsets obtained for a smaller number of parameters. This is due to the non-linear nature of the dependency, which requires the use of regression algorithms capable of handling surfaces with multiple local minima. Genetic algorithms have been shown to be particularly effective for this purpose (Davis 1991, for a comprehensive overview). We have used this iterative optimisation procedure not only to select the most discriminant tests defining a MCT but also to optimise the test itself, e.g. to determine the optimal incubation time of a substrate utilisation test. In figure 1 the use of entropy calculations is exemplified to evaluate the selection of 3 out of possible 4 tests (T1,2,3,4) by their capacity to discriminate two samples (A,B). The four possible solutions can be codified as binary vectors, e.g. Solution 2 comprises tests T1,2 and 4 (Fig.1b, Sol.2). The binary vectors are handled by genetic algorithms that search for the solution with the highest entropy (Fig.1c, Sol.3 and 4).

a)

Data:

	T1	T2	T3	T4
A	1	1	1	0
B	1	1	0	1

b)

Solutions:

	Sol1	Sol2	Sol3	Sol4
T1	1	1	1	0
T2	1	1	0	1
T3	1	0	1	1
T4	0	1	1	1

c)

Class frequency and entropy for each solutior

A	B	Sol1	Sol2	Sol3	Sol4
0	0	0	0	0	0
0	1	0	1/3	1/3	1/3
1	0	1/3	0	1/3	1/3
1	1	2/3	2/3	1/3	1/3
	S=	0.64	0.64	1.1	1.1

Figure 1 - Example of utilisation of entropy measurement to select 3 out of 4 possible binary tests to discriminate two categories, A and B. a) experimental test results; b) possible solutions expressed as a binary vector ("genetic" sequence format used by the genetic algorithm); c) Calculation of entropy levels associated with each solution (eq.1): the optimal combination of 3 test typing is shown to be ***T3AND T4AND(T1OR T2)*** = *Sol.3 or Sol.4*.

MCT interpretation

The information encoded in the MCT has to be classified according to associated process parameters of interest, i.e. extent of contamination, rate of bioremediation (Fig.2). As outlined in the introductory section, the association is non-linear and the underlying mechanism is usually poorly understood. Therefore, instead of aiming at proving explicit associations between parameters, statistical analysis will have to be preceded by the development of predictors capable of using information implicit in the experimental record.

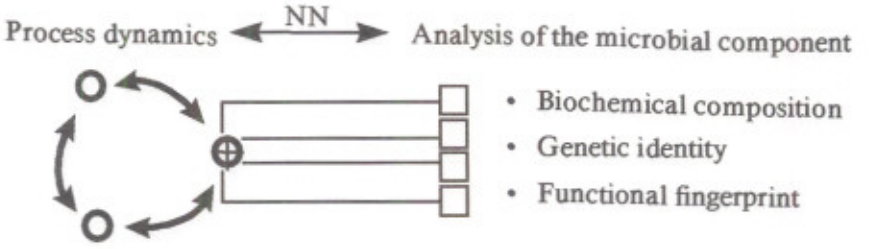

Figure 2 - Rationale for the use microbial composition typing (MCT) as a probing method. Process parameters are inferred from MCT by using artificial neural networks (NN).

Neural networks (NN) are an artificial intelligence technique that emulate the process of natural learning ("learning from experience", Hinton 1992). NN consist of a layered assembly of interconnected neurons, mimicking the biological counterpart. Each neuron is a processing unit that performs weighted sum of all inputs, x, and transfers the output, y, to the next layer. The transfer function is typically a sigmoidal curve. The matricial formulation for a layer of neurons is presented in equation 2, where x is the vector of n inputs, y is a vector of m outputs produced by m neurons; w is the mxn weight matrix; and wo is a vector of n scalars.

$$y = f(wo + w \cdot x) \quad , \quad f(\alpha) = 1/(1+e^{\alpha}) \qquad (2)$$

The learning process consists of simultaneously presenting to the NN typing profiles (e.g. Fig.1a) and classification vectors (e.g. Fig.1.c). The layer of neurons processing the typing profiles is the *input layer*; the layer whose outputs are compared with the classification vector is the *output layer*; the intermediate layers are known as *hidden layers*. The conventional learning algorithm backpropagates the classification errors to the values of the weight matrix and bias vector, a procedure akin to reinforced learning. For practical applications, other regression algorithms are often preferred (Haykin 1994 for a comprehensive reference on NN).

Three different NN architectures that have been used to interpret MCT profiles are hereby reviewed: standard feedforward, non-linear mapping and auto-associative. 1) Feedforward NN associate the MCT profile with the classification vector as described in Figure 3. The number of hidden nodes is allowed to change freely during the learning process. Feedforward NN of this type can theoretically emulate any transfer function, as suggested by the Komogorov theorem (Bishop 1995). In order to guarantee the generalisation of the NN solution, part of the experimental data is exclude from the learning process, and is used for validation. The appropriated number of hidden nodes (Fig. 3) and the extent of error backpropagation is determined by comparing the learning and validation errors (cross-validation, Masters 1995).

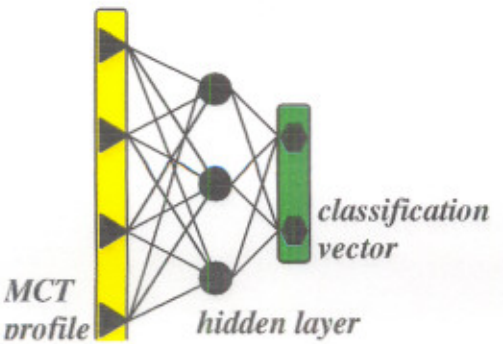

Figure 3 - Schematic architecture of a three layer Feedforward network to associate MCT profiles with classification vectors. Symbols correspond to neurons (eq. 1).

equally feasible to train several networks to extract different types of information from the same MCT profile, e.g. using PLFA profiles to infer both physiological status and identity (Almeida 1995, Pfiffner 1998)

2) Non-linear mapping uses a feed-forward NN with three hidden layers where the middle layer is constrained to the dimensions of the map (usually 2D, Fig. 4). A MCT profile analysed by this method is assigned to a position in the map that has unique properties with regard to the classification vector. As a consequence, a non-linear map classifies MCT profiles and also sorts process parameters according to their joint occurrence (Almeida 1998). In that report we have described application of non-linear mapping handling 6 parameters simultaneously: process time, contamination, inoculation with defined culture and the presence of three selected genes for degradation of hydrocarbon compounds in soil.

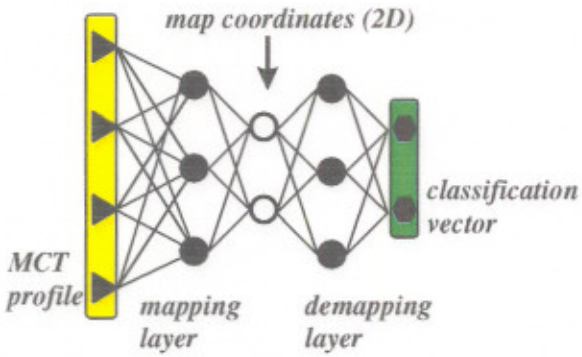

Figure 4 - Non-linear mapping of changes in the MCT profile that indicate changes in the membership in the categories contemplated by the classification vector.

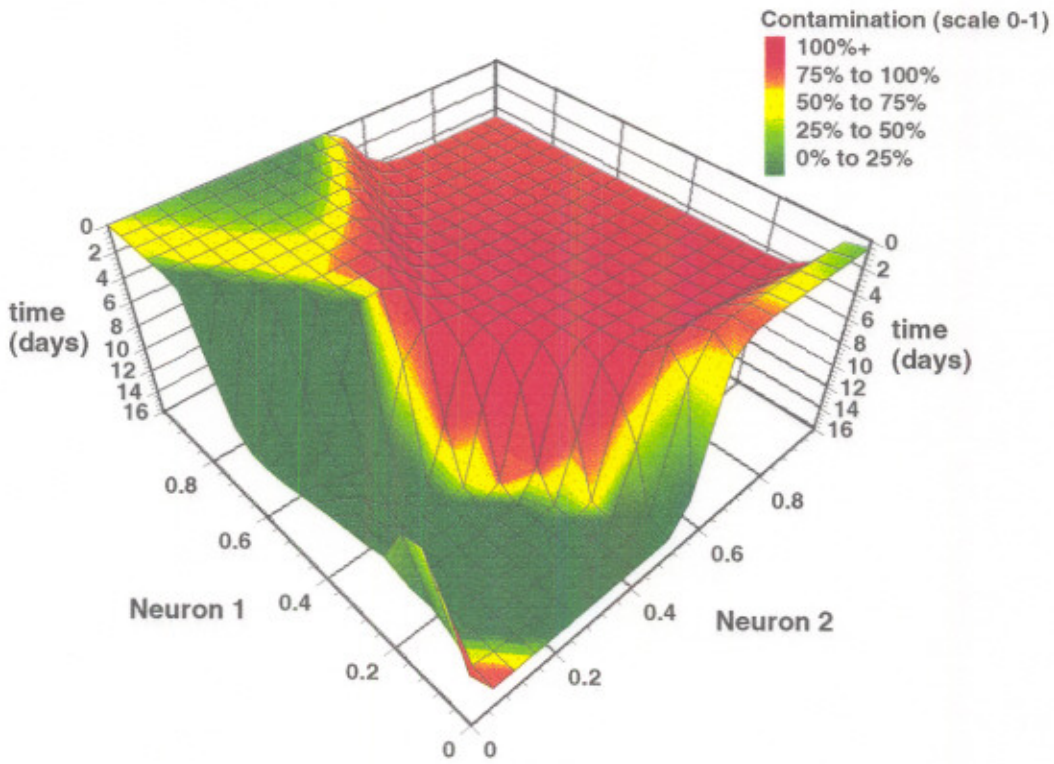

Figure 5 - Simulation of time course of soil hydrocarbon contamination levels by using the non-linear map reported in (Almeida et. al 1998). The two horizontal axis are the coordinates computed by the middle neurons (Fig.4); the vertical axis is the inverted time scale; the color-scale represents the contamination level. Time course of an individual sample can be inferred by following the steepest path starting from the assigned coordinates (see text).

The non-linear map developed was used to obtain the simulated contamination levels plotted in Figure 5. The vertical axis is an inverted time scale covering a period of 2 weeks. The colour-scale represents the hydrocarbon contamination level, from 100% (contaminated) to 0 (not contaminated). A soil sample whose microbial community had been typed by this MCT method and analysed by the non-linear NN mapping technique described will be assigned a position in the map. That position not only tells what is the contamination level but also suggests what is the future process course, which can be obtained by following the path with the steepest slope. Therefore, non-linear mapping of MCT profiles is as much a monitoring as a modelling tool. Another utilisation of NN hidden nodes to infer process course from MCT profiles was reported by Noble et al., 1997 where the pattern of hidden node outputs is subjected to cluster analysis.

The two achitechtures presented above were used to associate typing profile with process parameters. However, the relevant process parameter may not be available or may have been unreliably determined. The third architecture, 3) autoassociative NN, uses the vectors of MCT profiles as both primary input and final output (Fig. 6). The procedure is conceptually similar to dimensionality reduction by principal component analysis, except that the components are curves with a flexible shape instead of lines (Bishop 1995). The ordination of PLFA profiles by this method represents a higher proportion of experimental variance than does principal component analysis.

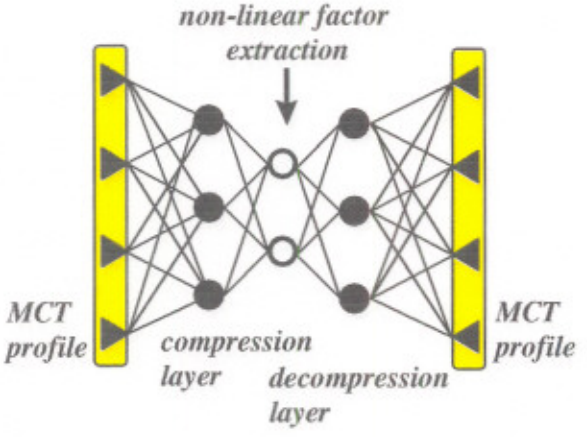

Figure 6 - Autoassociative architecture for non-linear factor extraction

Although the literature reports on analysis of MCT profiles mostly by standard multivariate statistical methods, neural networks have clear advantages. The recent statistical literature reflects this new awareness by increasingly reporting on neural computing techniques for cluster analysis, factor analysis and discriminant analysis (Cheng 1994).

Database management and integration

In order to optimise the analysis of MCT data, it is of critical importance to store all available data in a central relational database. Only then can the NN methods presented above, be used to explore the interdependencies that may be implicit in the experimental record. In addition, having a relational database as a central data repository enables the combination of multiple MCT techniques to develop a *committee* (Fig.7). Committee classifiers are particularly effective when the member classifiers have specialised competencies (Bishop 1995).

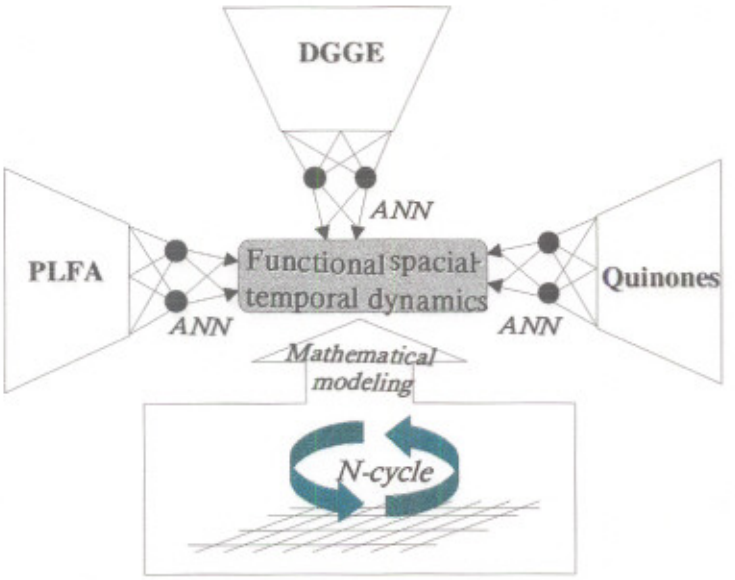

Figure 7 – Example of a combination of multiple MCT methods for a comprehensive description of microbial mediated nitrogen cycling (compare with Fig.1). A single relational database is regularly updated with field data, and concomitantly, constant calibration of neural network responses takes place.

Using MCT methods to monitor environmental processes mediated by microbes is often part of a broader management framework. Taking into account local heterogeneity and global dependencies brings forth the development of geographic information systems (GIS). The new context extends the role of microbial community typing (MCT) methods as a source of data for calibration of remote sensors (Fig. 8).

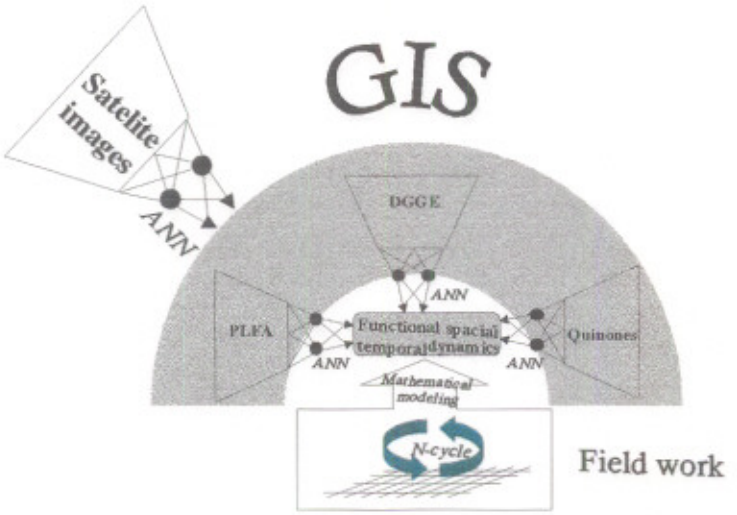

Figure 8 – Rationalization of using MCT methods (shaded half circle) for calibration of remote sensing within the context of geographic information systems (GIS).

Acknowledgements

This work has been supported by European Union Project IC18-CT0160 (DG12-SEDK), "Methodologies and design criteria for soil and water resource management and policy formulation in peri-urban farming systems in southern Africa"; by grant FMRH/BSAB/60/98 Fundação para a Ciência e Tecnologia /MCT; and by the Department of Energy DE-Fco2-96 Er62278, for Assessment Program Leadership in the Natural and Accelerated Bioremediation (NABIR) program, administered by J. Houghton.

References

Almeida, J.S., A.Sonesson, D.B.Ringelberg, D.C.White (1995) Application of artificial neural networks (ANN) to the detection of Mycobacterium tuberculosis, its antibiotic resistance and prediction of pathogenicity amongst Mycobacterium spp. based on signature lipid biomarkers. *Binnary-Computing in Microbiology*. **7**, 53-59.

Almeida, J.S., K. Leung, S.J. Macnaughton, C. Flemming, M. Wimpee, G. Davis, D.C.White (1998) Mapping changes in soil microbial community composition signaling bioremediation. *Bioremediation J*, **1**, 255-264.

Bakken, L.R. (1985) Separation and purification of bacteria from soil. *Appl. Environ. Microbiol.* **49**, 1188-1195.

Bishop, M.P. (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford, pp.137-140 Komogorov theorem; pp.310-319 factor analysis;

Cheng, B., D.M. Titterington (1994) Neural Networks: a review from a statistical perspective. *Statistical Science*, **9**, 2-54.

Davis, L. (1991) The Handbook of Genetic Algorithms. Van Nostrand Reingold, NY.

Di Gregorio, S. (1997) A Cellular Automata Model of Soil Bioremediation. *Complex Systems* **11**, 31-54.

Garland J.L. (1996) Analytical approaches to the characterization of samples of microbial communities using patterns of potential C source utilization. *Soil Biol. Chem.* **28**, 213-221.

Hastings, A., K. Higgins (1994) Persistence of Transients in Spacially Structured Models.. *Science* **263**, 1133-1136.

Haykin, S (1994) Neural Networks, a comprehensive foundation. Macmillan College Publishing Co., NY.

Hinton, G.E. (1992). How neural networks learn from experience. *Sci. American*, **267**, 144-151.

Kawaharasaki, M., T. Kanagawa, H. Tanaka, K. Nakamura (1998) Development of 16S rRNA targeted Oligonucleotide probe for detection of the phosphate accumulating bacterium Microlunatus phosphovorus in an enhanced biological phosphorous removal process. *Wat. Sci. Tech.* **37**, 4-5.

Kowalchuk, G. A., J. R. Stephen, W. De Boer, J. I. Prosser, T. M. Embley, and J. W. Woldendorp (1997). Analysis of ammonia-oxidizing bacteria of the beta subdivision of the class Proteobacteria in coastal sand dunes using denaturing gradient gel electrophoresis and sequencing of PCR amplified 16S rDNA fragments. *Appl. Environ. Microbiol.* **63**, 1489-1497

Liu W.T, T.L. Marsh, L.J. Forney (1998) Determination of the Microbial Diversity of Anaerobic-Aerobic Activated Sludge by a novel molecular biological technique. Wat. Sci. Tech. **37**, 4417-422.

Logan, N.A. (1994) Bacterial Systematics. Blackwell Scientific Pub., Oxford, pp.13-34, 47-61.

Macnaughton, S. J., T. L. Jenkins, M. H. Wimpee, M. R. Cormier, and D. C. White (1997) Rapid extraction of lipid biomarkers from pure culture and environmental samples using pressurized accelerated hot solvent extraction. *J. Microbial Methods* **31**, 19-27

Pahl-Wostl, C. (1995) The Dynamic Nature of Ecosystems. John Wiley & Sons, NY, USA. pp.44-87

Pfiffner, S.M, C.C. Brandt, J.C. Schryver, A.V. Palumbo, J.S. Almeida (1998) Using Artificial Neural Networks to Assess Microbial Communities. *Proceedings of the National Conference on Environment Science and Technology*, Greensboro, North Carolina State Agriculture and Technical University.

Pierce, J.R. (1980) An Introduction to Information Theory. Dover Pub., NY, pp.19-63 for historic overview of communication theory, pp.78-106 for measures of entropy.

Skinner, F.A., P.C.T. Jones, and J.E. Mollison (1952) A comparison of a direct and a plate counting technique for the quantitative estimation of soil microorganisms. *J. Gen. Microbiol.* **6**, 261-271.

Smalla K, Wachtendorf U, Heuer H, Liu WT, Forney L. (1998) Analysis of BIOLOG GN substrate utilization patterns by microbial communities. *Appl Environ. Microbiol.* **64**, (4) 1220-1225.

Stephen, J. R., G. A. Kowalchuk, M.-A. V. Bruns, A. E. McCaig, C. J. Phillips, T. M. Embley, and J. I. Prosser. (1998) Analysis of β-subgroup proteobacterial ammonia oxidizer populations in soil by denaturing gradient gel electrophoresis analysis and hierarchical phylogenetic probing. *Appl. Environ. Microbiol.* **64:**2958-2965

Voordouw, G. (1998) Reverse sample genome probing of microbial community dynamics. *ASM News* **64**, 627-633.

White, D.C., J.O. Stair and D.B. Ringelberg (1996) Quantitative comparisons of in situ microbial biodiversity by signature biomarker analysis. *J. Indust. Microbiol.* **17**, 185-196.

White, D.C., K. Leung, S.L. Maccnaughton, C. Flemming, M. Wimpee, and G. Davis (1997) Lipid/DNA biomarker analysis for assessment of *in situ* bioremediation effectiveness. In: *In Situ and On-Site Bioremediation*, **5**, 319-324, Battelle Press, Columbus Ohio.