CRANFIELD UNIVERSITY

INSTITUTE OF BIOSCIENCE AND TECHNOLOGY
CRANFIELD UNIVERSITY AT SILSOE

PhD THESIS

Academic year 2002-2003

S. M. O'FLAHERTY

MICROBIAL DIVERSITY IN CONTAMINATED SOIL

Supervisors: N Magan, P Hirsch

September 2002

This thesis is submitted to fulfil the requirements for the degree of Doctor of Philosophy.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

To Éadaoin.

Good things come to those who wait.

Abstract

Little is known about soil population dynamics: how microbial communities are affected by environmental stress; whether they respond adversely, or adjust to it; what the extent and duration of these adjustments might be; which genera of bacteria are the most reliable indicators of soil quality. Inevitably, it is beyond the scope of this thesis to resolve all these issues, but it nonetheless sets out to address specific objectives which should cumulatively enhance our total understanding of the soil microbial environment.

A three year study of abiotic environmental stress, using a distinctive multi-disciplinary methodology, examined how pseudomonad communities react to chronic metal pollution from sewage sludge. It deployed three profiling methods: population size, catabolic and genetic diversity, across three sequential sampling times, and processed large numbers of bacterial isolates to facilitate meaningful data analysis.

This process required innovative methodologies. Efficient analysis of numerous pseudomonad isolates necessitated semi-automation, by adapting ERIC profiling to run on Genescan, a specialist application for analysing sequence data, hitherto unused to fingerprint soil isolates. Specialist computer programmes were designed to compare multiple isolates. A database system was built to gather ERIC profiles and convert them to generate standardised data for subsequent analysis. The main body of this thesis focuses on the interpretation of genetic relationships, largely derived from ERIC profiles, for which a series of programs was built. Alongside them, the BIOLOG™ technique is deployed to profile catabolic diversity and thus compare the function of environmental samples at different sampling times. Population composition, a third dimension, was examined using conventional plate counts.

Initial findings that population diversity might be affected by metal stress were not, in the end, corroborated by the extended study. Both catabolic and population studies proved inconclusive, highlighting only subtle differences between metal-contaminated plots. Soil pseudomonads, in contrast to rhizobia, failed to prove sensitive indicators of low level metal contaminants.

Acknowledgements

The debts I owe in the completion of this thesis are many and varied. In the last 15 months before submission, the birth of my baby daughter brought with her an intriguing series of complications. Everyone mentioned below deserves my deepest gratitude, as players in that endgame, or else long distance runners; often both.

I would like to acknowledge Dr Penny Hirsch, who deserves special mention for her unstinting support, both professionally and personally, over the past four years. Professor Naresh Magan, of Cranfield University, has provided invaluable advice and assistance. I would also like to thank Darren Murray for providing both outstanding and patient statistical advice, and Kev Coleman for fixing my *endless* computer problems. They are both consummate professionals.

I would also like to thank my colleagues for their professional support and friendship. Those deserving special mention are: Tom Mendum, Ian Clark, Simon Atkins, Wendy Wilmer and Maureen Birdsey. Recognition should also go to all members of Penny Hirsch's Group, past and present, who lent their support in very practical ways over the years. In fact, to everyone who helped in any way, great or small, in making this work possible, I thank you.

My personal obligations, importantly, should not go unacknowledged. My family (especially Susan Abrook and Willie O'Flaherty) and close friends, provided the support which has made this submission possible. In particular, I would like to extend special thanks to my husband, Robin, who, in his turn, spent many hours patiently proof-reading drafts. His willingness, especially in the final stages, to take our new-born baby daughter for walks of staggering length and ingenuity gave me the space I needed to complete on time.

Lastly, and perhaps most significantly, I would like to thank my six-month old daughter herself. Molly Éadaoin is too young yet to understand her significance in providing me with the impetus, and the will, to finish up; on this basis, I would like to dedicate this thesis, and all of its joys and tribulations, to her.

List of Contents

Chapter One: Introduction

Chapter Two: General Methods

Chapter Three: **Population Size and Composition**

Chapter Four: Catabolic Diversity

Chapter Five: Genescan System Development

Chapter Six: Access Database Development

Chapter Seven: Fingerprint Data Analysis

Chapter Eight: Fingerprint Method Development

Chapter Nine: Genetic Diversity in FYM and Sewage

Chapter Ten: **Conclusions and Implications**

List of Figures

Chapter Eight

Chapter Nine

List of Tables

Chapter Five

Chapter Six

Chapter Seven

Chapter Eight

Chapter Nine

Chapter One:
Introduction

1.1 Introduction

Microbial diversity is increasingly acknowledged as fundamentally important to ecosystems, including soil. Large and diverse populations of microbes are important indicators of soil health. Many carry out essential processes, like nutrient cycling. Others are critically important for healthy plant growth (Pankhurst *et al.,* 1995). This 'dependency chain', which extends upwards from healthy microbial communities, has a simple and compelling logic. Healthy soil is not only important as a means of supporting crop production, but is also an essential natural resource, which should be managed for future generations. If it is not, then the consequences are likely to be both widespread and adverse; even the loss of specific microbes could have direct and damaging economic consequences for agriculture.

We are right to be concerned. Although it is currently difficult to predict the exact consequences a loss in diversity may entail, there is mounting concern over the long-term environmental effects of industrialisation and urbanisation, which have gathered in pace and intensity (and without interruption) since the nineteenth century. Some measure of soil quality is necessary to determine whether soil is being degraded prior to irreversible larger scale damage; diversity, we believe, deserves to be considered a potentially appropriate indicator of it.

Nonetheless, we should acknowledge the limitations of what we know of microbial communities, and how to measure them; in contrast, our current understanding of plant and crop ecosystems is extensive. One of the central reasons for the difference is that soil microbial ecology, a vast and complex field of enquiry, is only superficially understood. Yet despite this comparative lack of knowledge, new techniques and approaches, some of which are discussed in detail in this thesis, should cumulatively enhance our ability to understand the soil.

That will take time. In the here and now, however, even without fully understanding the nature and extent of environmental damage caused by reduced diversity, it nonetheless

seems prudent to preserve the *status quo*. Diversity, both functional and genetic, is not only important in preserving ecosystems; it is also crucially important to the agricultural economies built on them. Moreover, healthy soil is not just about sustaining a chain of *dependency*; it is also a potentially invaluable source for future technological advance (new methods, new products; Torsvik *et al.,* 1996).

We therefore advocate a policy of 'containment', ie that soil microbial diversity should be preserved in its present condition, as insurance for the future, if nothing else, while we pursue concurrent research to decode its full meaning (Bengtsson 1998).

1.2 Diversity

Diversity is an expression of the variety of living things, at the genetic, species and ecosystem levels (Pankhurst *et al.,* 1998). Known microbial communities really do constitute the tip of the iceberg. The International Journal of Systematic Bacteriology listed 3800 characterised species in 1993, from an estimated number of between 300,000 and 1,000,000 species of bacteria, of which culturable bacteria represent only a small fraction. Hawksworth (1991) reported 1.5 million fungal species, representing less than 5% of extant species in fungal group, while Hammond (1992) has estimated that there are 5,000,000 extant viral species. Estimates of diversity range from five million to 1.4 billion species, while a study by Torsvik revealed that as many as 5,000 species of bacteria could exist in a single gram of soil (Torsvik *et al*, 1996). Soil micro-organisms reportedly constitute less than 0.5% of total soil mass (Tate *et al.*, 1995), but they are essential in nutrient cycling and exert a major influence on soil fertility (Figure 1.1).

From the vast number, then, of microbial species in the soil, little is known about the majority of them, either as organisms, or as bacteria performing a specific ecological role. Whether the *genetic diversity* of soil microbes, or the diversity of microbial *functions* is the most important indicator of soil quality, remains unclear. In fact, the relative abundance of different components within a community, and the interactions occurring between them, may be *equally* significant indicators of soil ecosystem diversity (Pankhurst *et al.*,1995). However, the identification of species involved in key soil processes allows us to study their function within the ecosystem. Quantifying these

effects is the first step in the sustainable management of soil quality (Bengtsson 1998), but no single process in current use reflects overall soil health.

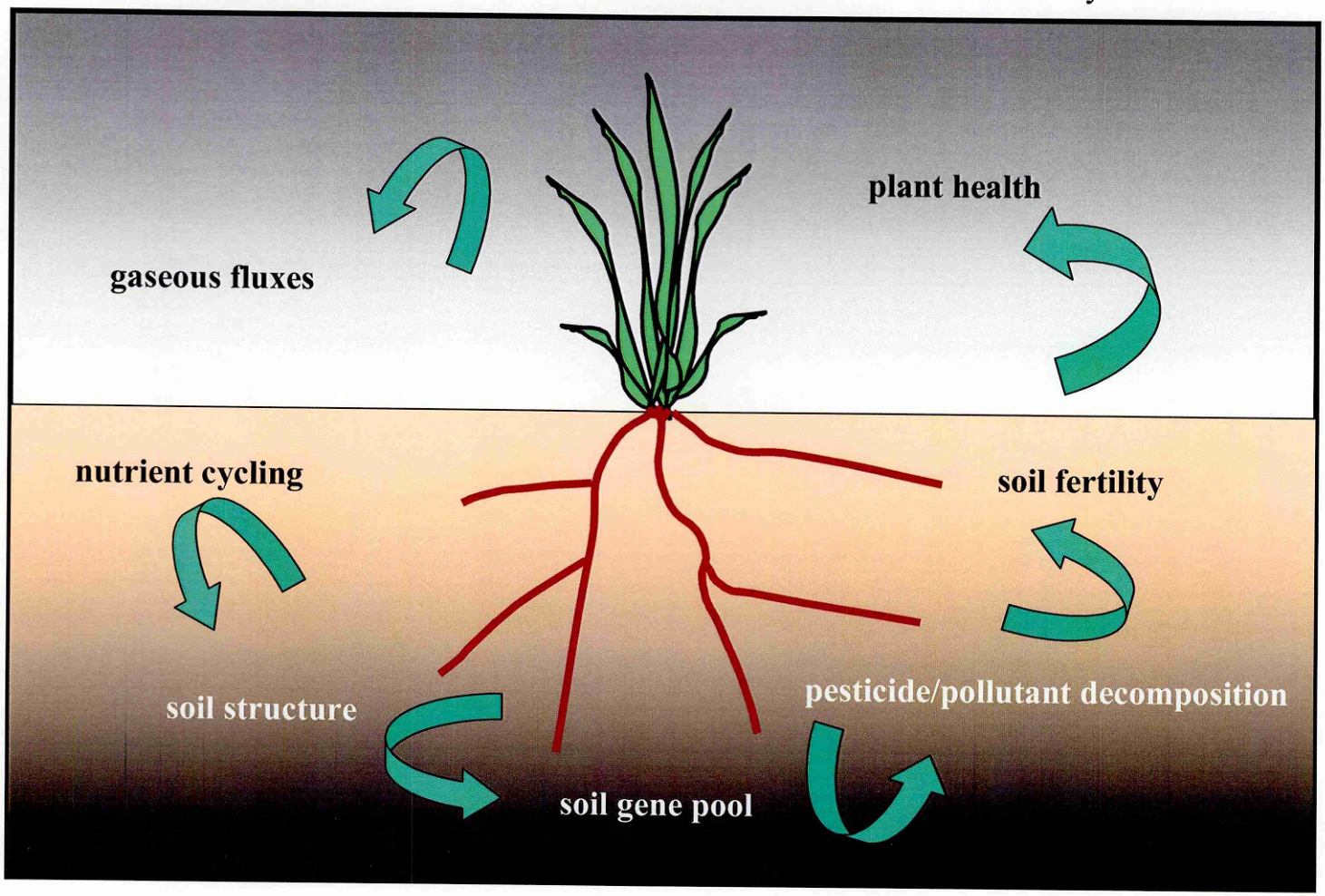
Diversity can be determined using phenotypic or genotypic approaches, and can be expressed as either species-or group-diversity; physiological diversity; or genetic diversity. (Torsvik *et al.,* 1996). The simple enumeration of microbial species, is a traditional, if rather one-dimensional, assessment of diversity (Pankhurst *et al.,* 1998). Diversity is calculated by means of a mathematical index, from a series of three commonly used indices. **Species Diversity** is an index expressing the ratio between the number of species, and the number of individuals, in a collection. It is a composite measure of richness and evenness, two additional statistical properties used to quantify species diversity. The former, **Species Richness,** is the number of species in a sample, whereas **Species Evenness** is the relative distribution of individual isolates within that sample.

These indices are mathematical calculations, originally used to measure *macro-organisms*. Now these measurements are routinely applied to microenvironments, an accurate description of each diversity measure is important. Different indices are chosen according to the type of analysis required, and classification systems used, thereby allowing for easier interpretation of experimental results, especially when comparing different microenvironmental communities.

1.2.1 Classification

To study microbial populations in the soil ecosystem, a means of assessing population structure is needed. Methods of defining taxonomic classification are well-established (e.g. molecular systematics) and these can be deployed to examine population structures (Priest *et al.,* 1993). This thesis does not propose the *systematic* classification of bacterial groups, but to investigate the basic structure of a soil community, similar analytical processes are needed. However, when applying these methods, care must be

Figure 1.1 Structural interactions from and within the soil microbial community



taken not to over-interpret results, since they neither reflect taxonomic structures, nor phylogenetic relationships. Nevertheless, basic classification methods are necessary to understand the processes which are applied in systematic studies. There are three different methods for the classification of microorganisms (Priest *et al.,* 1993), which are:

1.2.1.1 Special purpose classification

This is a classical method, which depends on precise information on the functionality of microorganisms. Special purpose classification is 'artificial' in the sense that it does not display relationships between organisms, but concentrates instead on one or two important characteristics which differ. This type of classification, used routinely in clinical microbiology, identifies differences in pathogenic properties, for example distinguishing between *Shigella dysenteriae* and *Escherichia coli.* This phenotypic method of classification is too specific for our purposes and is thus unsuitable for environmental studies.

1.2.1.2 Natural (phenetic) classification

This is a general-purpose classification method which displays relationships (ie affinities) based on complete organisms (genotype and phenotype) as they presently exist, based on 'higher numbers of common features', with no reference to the evolutionary pathway, or ancestry of the organisms. Functional assays such as BIOLOG GN™, which compare carbon substrate utilisation, are an appropriate method of measuring diversity in the soil ecosystem, providing catabolic profiles of fast growing heterotrophic bacteria in soils.

1.2.1.3 Phylogenetic classification

This is based on the phylogenetic (genealogical) relationship between organisms and attempts to trace the evolutionary pathways which have given rise to the organisms, with the classification exactly reflecting the line of ancestry. It relies on genetic relationships and is the most relevant way of looking at genetic diversity in the soil microbial population.

The ordering of environmental isolates into a systematic form is generally accomplished via phylogenetic classification, but it should also be applied in conjunction with natural classification (Priest *et al.,* 1996). The latter involves the generation of a large database of organisms which are then grouped into 'taxo-species' (i.e. clusters) on the basis of shared similarities, hence generating a probability matrix, which is then used to produce phylogenetic trees.

1.2.1.4 Phylogenetic trees

The relationship between different 'clusters of isolates' in a population can be expressed using mathematical terms. Numerous equations are used to calculate these relationships, but a distance coefficient -the 'Euclidean distance'- is generally used (Priest *et al.,* 1996). This is an arbitrary measurement which can be used successfully for binary data. There is a suite of phylogenetic computer programs (Phylip™) available to produce phylogenetic trees (Felestein 1996), while hierarchical procedures with cluster analysis (UPGMA) are generally used to perform ranked classifications. Although defined as phylogenetic, the 'Euclidean distance' is reported to be more phenetic (ie, natural) (Priest *et al.,* 1996), and can therefore be used to analyse environmental population structures. This contrasts sharply with more traditional bacterial taxonomy, where taxa are defined and recognised using many equally weighted features, namely behavioural, morphological and dye-binding properties. Phylogenetic analysis programs can be deployed to produce 'cluster' information on environmental isolates, but should not be used to infer *genetic* relationships.

1.2.2 Microbial diversity studies

Soil microbial diversity can, and should, be studied at three different levels, each important in its own right, and cumulatively able to supply a broader perspective on micro-ecosystems. Indeed, by studying different levels of diversity in conjunction, new insights into the community structure may be generated. These three levels comprise:

1. Taxonomic: a simple enumeration, which profiles distribution, size and diversity of specific populations;
2. Functional: catabolic diversity, which studies *functional* components of the ecosystem;
3. Genetic: genetic variation (diversity) of populations.

1.2.2.1 Population composition and size diversity

To appreciate the consequences of biological activity, the living component of soil must be considered (Tate 1988). To this end, it is appropriate to evaluate culturable microbial sub-communities. Because the relative size of various populations is determined by their ability to compete with other organisms, measurement of the latter is, in many soil ecological studies, a necessary initial step. Population composition, size and diversity is collectively a simple measure which provides a one-dimensional profile of soil population, against variables of different stresses within the soil environment. It is generally accepted as a quantitative analysis (or 'headcount') of selected culturable sub-populations. The data obtained can reasonably be criticised for lacking the breadth and detail necessary to determine the full effects of soil population stresses. Nonetheless, it is invaluable in gathering preliminary information on properties of the microbial community within the soil ecosystem.

1.2.2.2 Catabolic diversity

Functional diversity may represent a number of different functional or ecological groups (Torsvik *et al.*,1996). This functional diversity has recently been expressed as a measurement of the ability of the microbial communities to utilise a range of carbon sources (Garland 1996; Degens and Harris 1997; Reber 1992; Wenderoth *et al.*, 1999). The analysis of soil community functioning based on utilisation of substrates, using a commercially available system- BIOLOG™, is one of these measurements used to study catabolic diversity (Garland and Mills, 1991; Zak *et al.*, 1994). The BIOLOG™ system allows for a rapid, community level approach for assessing patterns of carbon utilisation, by mixed microbial samples. This community-level physiological profiling, or catabolic diversity, is effective at distinguishing spatial and temporal changes in microbial communities (Garland 1997, Garland and Mills, 1991).

The BIOLOG GN™ System, developed and introduced in 1989, uses a GN Microplate panel containing 95 different carbon substrates for the identification of a wide range of Gram-negative species. This commercially available identification kit was originally targeted at the identification of microbes affecting humans and was designed to be used as a microbial identification system. It was soon realised, however, that another

application was feasible and it was extended to measure functional diversity in soil communities, by metabolic profiling (Garland 1986; Pankhurst 1998).

Each GN Microplate contains 95 carbon utilisation test wells, with each well containing nutrients (a carbon source), salts, a small amount of peptone and redox dye: tetrazolium violet (Bochner 1989, Winding 1994). A control well, for comparison, contains no carbon source. Utilisation of the carbon sources is detected as an increase in respiration by living cells, which leads to an irreversible reduction in the tetrazolium dye. Hence, as the bacteria begin to metabolise, a purple colour is formed in the well, yielding a pattern of purple wells of differing intensities, which constitutes a 'metabolic fingerprint' of the capacities of the inoculated soil organisms. Another new BIOLOG GN™ system, EcoPlates™, was introduced for ecological analyses after the 1996 SUBMECO conference in Austria (Insam and Rangger 1997). At a select meeting, ecological statisticians decided that the 95 well plates produced too many variables and insufficient replication for satisfactory statistical analysis. EcoPlates™ Plates were therefore proposed, containing the same chemistry, but with a plate containing three sets of 31 carbon sources, alongside one reference well. Many of the 31 carbon sources, also present in the 95 well plates, were chosen because they highlight generally important metabolic traits in soil microbes, and (some) are representative of carbon sources found in soil.

Carbon substrate utilisation is relevant both to the ability of soil microbes to degrade crop residues, and to their ability to proliferate in the rhizosphere, where organisms compete to utilise carbon-rich root exudates. The BIOLOG GN™ Systems make it easier for environmental and ecological microbiologists to investigate bacterial populations rather that individual isolates, only a minority of which can be present in culture collections. They therefore facilitate the description of differences in functional diversity (metabolic profiling) among communities of soil bacteria. The system has been used to characterise soil microbial communities and provide a indication of how heterotrophic sub-populations (i.e. the active bacterial population) are composed (Zak *et al.*, 1994; Garland and Mills, 1991; O'Flaherty *et al.,* 1998).

1.2.2.3 Genetic diversity

Recent developments and applications of molecular approaches to microbial ecology have led to significant advances in the area of microbial identification and classification giving rise to a novel discipline: Molecular Microbial Ecology (de Bruijn 1992). There are now numerous molecular techniques which enable new approaches to microbial ecology and increase our understanding of environmental genetic diversity. These techniques can be divided into two categories:

1. those which do not rely on bacterial culture, but rather investigate DNA extracted directly from environmental samples;
2. methods involving culture of bacteria prior to genetic analysis.

1.2.3 Advances in molecular microbial ecology

Arguably, one of the most important advances as regards molecular biology in the last century, because it permits analysis of the molecular structure of organisms with unprecedented speed and accuracy, is the discovery and development of *Taq* polymerase and the polymerase chain reaction (PCR reaction) (Saiki *et al.*, 1988).

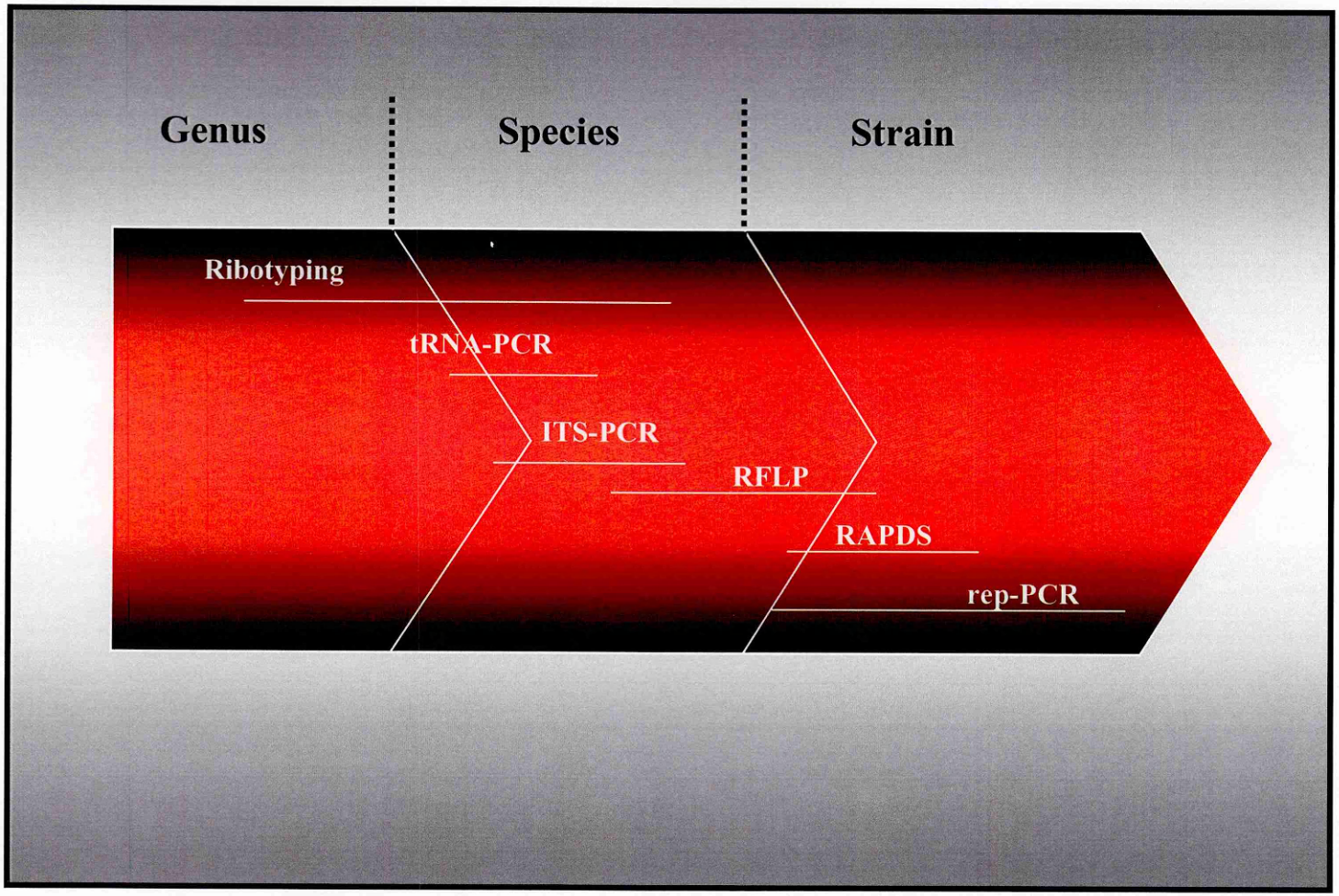
This crucially important technique requires some knowledge of DNA sequences, as specific or general DNA primers are required to bind to recognition sites either side of a region of DNA. These primers facilitate the exponential amplification of that region by *Taq* polymerase. There are a number of molecular methods which target specific sequences and facilitate the detection or identification of microorganisms. Such sequences may be identified via restriction fragment length polymorphism (RFLP-PCR), hybridization to a specific probe, or cloning, sequencing and analysis. The nucleic acid-based protocols often closely reflect phylogenetic relationships, which is useful for arranging strains into coherent groups. Therefore, these protocols offer an appropriate way of assessing possible changes in genetic diversity in a soil microbial structural population. The ribosomal genes are particularly useful: the 16S rRNA gene contains sequences conserved at the species, genus, family (or higher) levels, enabling different levels of discrimination depending on the method used.

DNA protocols employed for the routine determination of genera and species of microbes have been reviewed (Stackebrandt *et al.,* 1991, Akkermans *et al.,* 1997) The most relevant are:

1) the digestion of total genomic DNA, using restriction endonucleases which cut frequently (restriction enzyme analysis, REA) analysed by gel electrophoresis; or that cut infrequently, analysed using pulsed field gel electrophoresis (PFGE) or field inversion gel electrophoresis (FIGE);
2) plasmid profiling to examine extrachromosomal DNA;
3) restriction fragment length polymorphism (RFLP) analyses, where genomic digests are separated by gel electrophoresis, blotted and probed with an appropriate sequence;
4) 16S ribosomal gene analysis (e.g. ARDRA, RFLP, sequencing) also referred to as ribotyping;
5) Random Amplified Polymorphic (RAPDS) DNA and Repetitive Extragenic Palindromic (rep-PCR), similar PCR techniques producing arbitrary genomic fingerprints.
6) ITS-PCR: where several bacterial genes with related functions are arranged in clusters with intergenics spacers (IGS) of variable lengths. (The IGS with the ribosomal operon are transcribed and for this reason are generally called intergenic transcribed spacer (ITS) and the non-transcribed segments separating the repeats are referred to as IGS. The IGS and ITS regions are assumed to be less subjected to selection pressure and thus should accumulate more mutations than genes, and can therefore discriminate between closely related strains);
7) tRNA-PCR is a RNA fingerprint technique using the total cytoplasmic tRNA pool to provide a fingerprint specific to individual bacterial species, or genera.

These techniques are frequently used to provide information on the species-subspecies-strain level (de Bruijn 1992). Their relative ability to differentiate between a genus, species, or strain is shown in Figure 1.2.

Figure 1.2: Microbial classification methods based on amplification of DNA by the polymerase chain reaction.



1.2.3.1 Polymerase chain reaction (PCR)

There are now variations of the PCR technique which permit the typing of bacteria at strain level. These techniques use oligonucleotide primer sets constructed in the absence of specific sequence information, known as randomly amplified polymorphism DNA–PCR (RAPD-PCR) or as arbitrarily primed PCR (AP-PCR). These methods have been applied to many groups of organisms and have great potential to assist in understanding microbial community structures and interactions in an ecosystem (Rosado *et al.,* 1997). Two broadly-defined classes of PCR-based protocols are used in molecular ecology:

1. Amplification of a known fragment of DNA

This approach can be used when DNA sequence information about a particular gene or region is available. The amplified product itself, or the digest pattern of the product, is diagnostic of a genus, species or strain. A common use of this approach is with the ribosomal (rRNA) operon. Amplification of fragments are sometimes followed by DNA sequence or restriction analysis (ARDRA) of the amplified PCR product.

2. Amplification of fragments flanked by short DNA sequences scattered throughout the genome

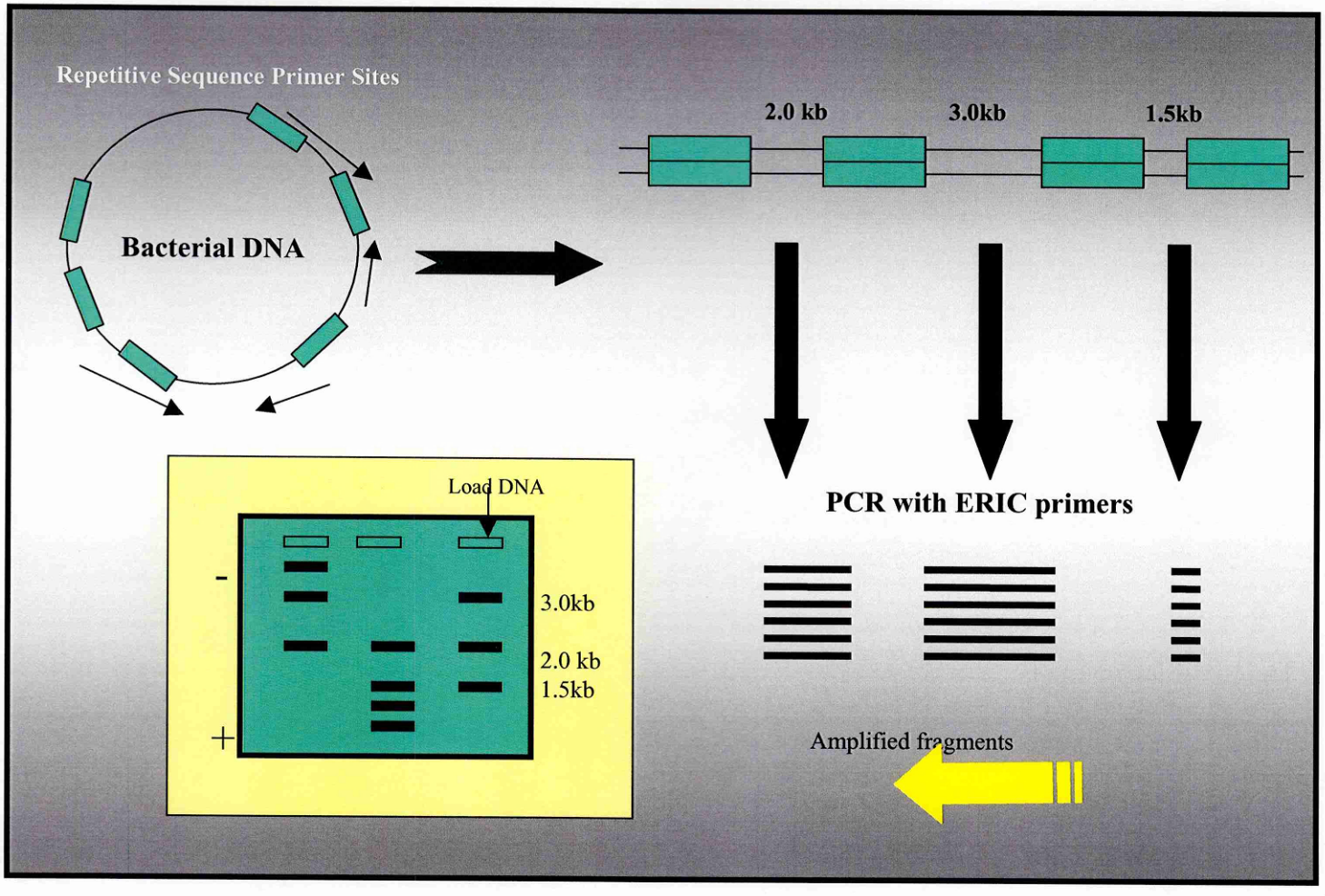
When these fragments are size-fractionated on a gel, the resulting banding patterns yield species or strain-specific 'fingerprints'. Two general protocols have been developed and applied to micro-organisms:

1) employs short primers of arbitrary sequence (RAPDs);

2) uses longer arbitrary primers, combined with low annealing temperatures (AP-PCR) to amplify DNA fragments of variable lengths.

The RAPDs protocol has been successfully applied to rhizosphere bacteria (Bruijn *et al.,*1992, 1995), hence this presents an ideal PCR-based technique which may be applied in this phylogenetic study (Figure 1.3).

Figure 1.3 Schematic representation of rep-PCR mediated amplification of genomic DNA sequences between repetitive elements



1.2.3.2 Ribosomal rRNA

Ribosomes, vital to all organisms, are a key element of the protein-synthesising machinery. The molecule consists of proteins and three small subunit RNA molecules, rRNAs, which vary in size. In prokaryotes they are: 5S rRNA, 16S rRNA and 23S rRNA. Eukaryotes have larger subunits. These rRNAs constitute a significant component of cellular mass, and are easily recovered from most organisms. They are highly conserved throughout evolution and this is seen at the nucleotide sequence level (Woese *et al.*, 1983).

The genes for the ribosomal subunits are conserved, but the spacer regions are not constrained by a requirement to retain a function. So, they are observed to evolve more rapidly than the coding regions. Therefore, the 16S rRNA subunit is generally accepted as the best target for studying phylogenetic relationships because it is present in all bacteria, is functionally constant and is composed of highly conserved, as well as more variable, domains (Figure 1.4).

The full sequences of the 16S rRNA from many bacteria and organelles are known, just as the 18S rRNA equivalent from many eukayotic groups have been sequenced. The 16S rRNAs provide precise statistical determination of evolutionary affiliation over both short and long range evolutionary spans (close and distant relatives; Pace *et al.*, 1986). Nevertheless, questions have been raised about 16S rRNA sequence data and whether it is sufficient to guarantee species identity (Fox *et al.*, 1992). But 16S rDNA sequences are currently considered of definitive importance to the phylogenetic dimension of bacterial taxonomy and need to be included in studies on the genetic relationships between bacterial isolates.

Figure 1.4. Diagram of 16S rRNA structure

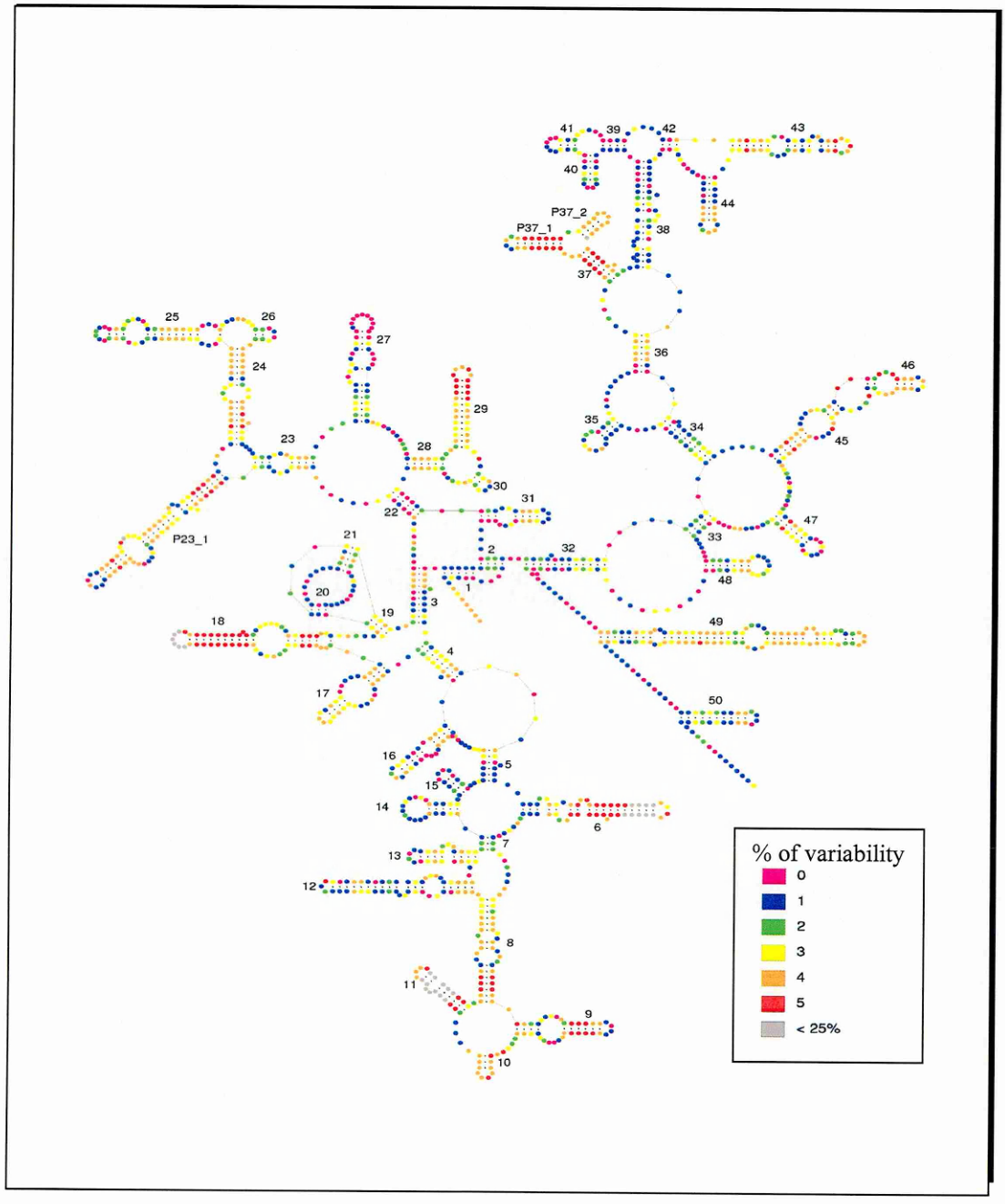


Table 1.1 Summary of primers for PCR amplification of eubacterial 16S rDNA

16S Universal Primers

Primers	Sequence (5'-3')	Designed for
fD1	AGA GTT TGAT CCT GGC TCA G	Most Eubacteria
fD2	AGA GTT TGA TCA TGG CTC AG	Enterics and relatives
fD3	AGA GTT TGA TCC TGG CTT AG	*Borrelia spirochetes*
fD4	AGA ATT TGA TCT TGG TTC AG	Chlamydiae
rD1	AAG GAG GTA ATC CAG CC	Many Eubacteria
rP1	ACG GTT ACC TTG TTA CGA CTT	Enterics (most eubacteria)
rP2	ACG GCT ACC TTG TTA CGA CTT	Most eubacteria
rP3	ACG GAT ACC TTG TTA CGA CTT	Fusobacteria (most eubacteria)

16S Pseudomonad Specific Primers

FS- for	GGT CTG AGA GGA TGA TCA GT	Most Pseudomonads
FS- rev	TTA GCT CCA CCT CGC GGC	Most Pseudomonads

The resolution of this method is determined primarily by the target of the PCR, i.e. if a spacer region is included, a higher discriminatory power will result. Primers of under 20 bases are used in arbitrarily primed PCR. A selection of primers is listed in Table 1.1 (Weiseburg *et al.*, 1991). The primers fD1 and rD1 were selected as a broad range universal primer set which would be successful with environmental isolates.

The primers FS-for and FS-rev were selected as highly selective primers specifically for the genus *Pseudomonas*. This highly selective PCR can be used as a rapid diagnostic tool for *Pseudomonas* genus identification (Widmer *et al.*, 1998).The *Pseudomonas* Selective PCR, in conjunction with RFLP analysis can be used to compare *Pseudomonas* population structures from a variety of ecosystems and provide insight into the occurrence, potential roles and possible unidentified subgroups of this genus in different ecosystems.

These PCR amplified rDNA products, when restricted, generate (mostly) species-specific patterns, which reflect the conserved character of the rRNA genes (Fox *et al.*, 1992).

1.2.3.3 RAPDs

RAPDs is a family of short intergenic repeated sequences. Several genomic fingerprinting methods are based on the amplification of these repetitive sequences (rep elements), within the bacterial genome. A number of repeat elements is known, but the most commonly reported repeat sequences for micro-organisms are:

- REP (repetitive enterobacterial palindromic)
- ERIC (enterobacterial repetitive intergenic consensus)
- BOX elements, repetitive sections of DNA (Versalvice *et al., 1991*).

REP and ERIC sequences, containing highly conserved central inverted repeats, can be dived into two classes that do not share significant homology (de Bruijn 1992). All these elements contain highly conserved palindromic inverted repeat sequences. Their actual function remains an enigma; no single function explains their DNA sequence conservation and/or ubiquitous distribution. Accordingly the suggestion has been made that these repeated elements represent "selfish" DNA sequences, which are maintained and propagated via gene conservation (de Bruijn *et al.,* 1995).

A number of methods have been developed to classify bacteria on the basis of their genomic DNA patterns. By using the PCR reaction (rep-PCR) and DNA primers, corresponding to the inverted repeats of naturally-occurring interspersed repetitive elements (REP, ERIC, BOX) in bacteria, a pattern of amplified fragments of different sizes can be obtained, corresponding to sites where the repeat element occurs (de Bruijn 1992). Hence the bacterial genome can generate a characteristic banding pattern on a gel ('fingerprint'). The intergenic repeat sequences used in this project are BOX and ERIC.

1.2.3.4 BOX intergenic repeat sequences

The discovery of the highly conserved DNA sequences located within intergenic regions of the chromosome of the Gram positive *Streptococcus pneumoniae* was first reported in 1992 (Bernard *et al.,* 1992). The highly conserved DNA elements are located within intergenic regions of the chromosome of *S. pneumoniae;* the genome contains about 25 of these elements called BOX (Bernard. *et al.,* 1992). The BOX sequences were the first group of highly conserved repetitive DNA elements found in Gram positive bacteria. The BOX sequences are thought to maintain themselves as

'selfish' DNA, and a possible functional role as regulatory elements would explain their persistence and sequence conservation in *S. pneumoniae* (Non-coding repetitive DNA is not likely in *S. pneumoniae,* given its small genome size compared with *E coli*). It is speculated that BOX represents a case of evolution of a repetitive DNA element recruited for a specific function, within the genome of *S. pneumoniae* (Bernard. *et al.,* 1992). The DNA Sequence for the BOX Primer is as follows:

5' CTA CGG CAA GGC GAT GAC G 3'

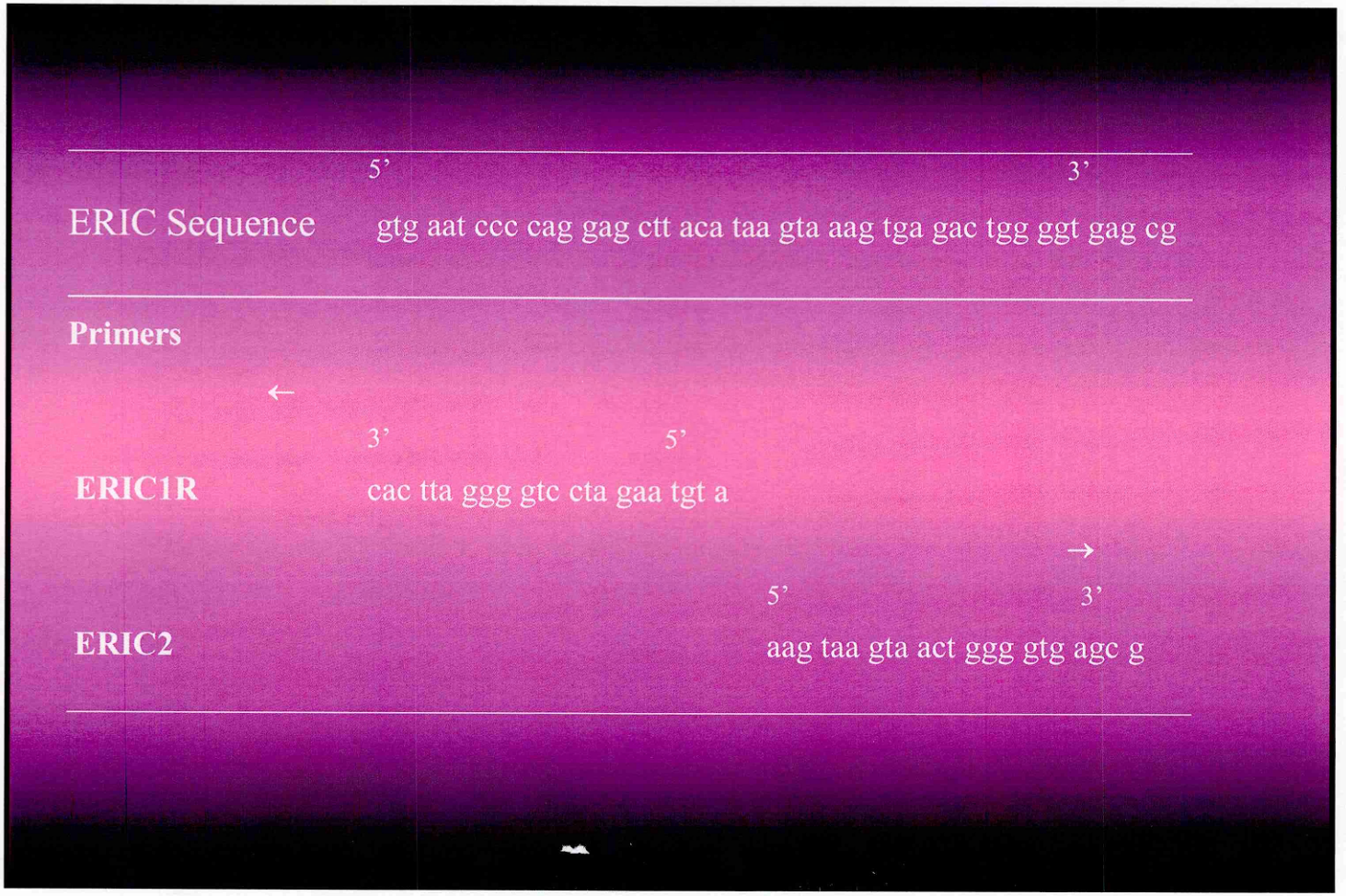
The BOX elements are unrelated to the two most thoroughly documented families of repetitive DNA sequences present in enterobacteria (REP and ERIC). Although the nucleotide sequences are entirely different, several characteristics of BOX are similar to those found in ERIC and REP elements. Amplification with BOX primers produces distinctive fingerprint patterns which have been used to determine phylogenetic relationships between bacterial isolates (Versalovic *et al.,* 1991). The BOX fingerprint method works on the same principle as ERIC, using recognised short, highly conserved palindromic inverted repeat DNA sequences within the bacterial genome. The amplification with BOX in a wide range of bacterial species results in a reproducible genetic fingerprint, which can be used for identification and to imply phylogenetic relationships (Versalovic *et al.,* 1995).

1.2.3.5 ERIC intergenic repeat sequences

Originally found in enterobacteria, this sequence has since been located in most bacteria and also some fungi. It has been found to be a helpful tool for studying relationships in families of Gram-negative soil bacteria (Figure 1.5).

The ERIC Fingerprinting method provides a very powerful tool to study diversity within a species, at a strain level, by analysing specific fingerprints generated from distinct genomes (Rosado *et al.,* 1997). Because it uses arbitrary sequences, it is mainly used for phylogenetic studies within well-defined genera, or single species. Specific DNA fingerprints have been generated from a wide variety of bacteria and used to

Figure 1.5 ERIC consensus sequences



determine phylogenetic relationships between microbial isolates and to study their diversity in natural ecosystems (Figure 1.3; de Bruijn, 1992). Preliminary work on pesticide-treated soils at Rothamsted employing this approach has identified a genetic shift in the pseudomonad population (Nicholson *et al.*, 1997), indicating that this technique is suitable for bacteria isolated from agricultural soils.

The ERIC Fingerprinting method permits a large number of different soils, and their respective isolates, to be studied and individual strains detected. Phylogenetic analysis can then be utilised to allow the formation of related groups derived from the unknown bacterial isolates. This permits observation of population structures in agricultural soils, allowing us to ascertain more effectively the impacts of environmental stresses.

It has been reported that a limited number of organisms and groups of organisms seem to control the critical processes necessary for ecosystem functioning: 'keystone process species' (Bengtsson 1998). These process species are vital for soil quality, but the species involved in such processes need to be identified. Because to date, no single key species has been put forward, I must postulate a potential keystone bacterial group and assess the microbial diversity within it.

1.3 Defining a key species

The 'Rhizosphere' is defined as the zone in which the microflora is influenced by plant roots (Rovira *et al.*,1965). The ecology of the plant-root system depends on the biology of the particular microbial community that the plant root system will support (Lynch, 1986). The distribution of different microbes in these communities may have negative and/or positive influences on rate of plant growth and subsequently on the yield of the crop. The bacterial genera *Pseudomonas, Bacillus* and *Clostridium* are reported to affect the rate of development and yields of a wide range of plants (Rovira *et al.,* 1965). They will, therefore, influence the general rhizosphere population indirectly via plant responses. Specific pseudomonad strains rapidly colonise roots of potatoes, sugar beet and radish, and are reported to cause statistically significant increases in yields in field trials (Lynch, 1986). Indeed, the direct effect of both *P. fluorescens* and *P. putida* has been strongly associated with a significant increase in plant yields. Another effect of

bacterial populations on plant-microbe interactions is via the Plant Growth Promoting Rhizobacteria (**PGPR**). **PGPR** enhance plant growth and increase crop yields, probably due to phytohormone production. The effect of PGPR is well documented and the colonisation of the rhizoplane on treated seeds or vegetative progating materials is termed bacterisation (Goto, 1986). Most PGPR so far described belong to the genera *Pseudomonas*, *Bacillus*, *Serratia*, *Arthrobacter* and *Streptomyces*.

A well documented aspect of microbe-plant interactions is the study of plant diseases. It is widely reported that the rhizosphere microflora affect the epidemiology of plant pathogens and the ability of some soil microbes to inhibit (directly or indirectly) the invasion of plant tissues by a plant pathogen. This effect can be mediated via direct competition between pathogen and root microbes; antagonism of the pathogen by root microflora; or alteration of root exudate diffusion into the soil environment, in a manner which interferes with the chemotaxis of the pathogen to the root (Tate, 1995). Soils in which plants are not susceptible to diseases, even though soil-borne pathogens are present in the soil, are termed 'Disease-suppressive soils'.

1.3.1 Disease-suppressive soils

An example of the natural suppression of a plant pathogen occurs with 'Take-All' (*Gaeumannomyces graminis* var *tritici*) which is recognised as a major pathogen of wheat, barley and related species. When this fungus attacks a healthy plant, it infests the root system causing darkening. The disease is only visible above ground when plants are killed and 'white heads', the fruiting bodies of the fungus, are formed on the plants (Lynch and Hobbie, 1988). However this disease can be controlled using effective land management practice. Control can be established by crop rotation, i.e. using a non-susceptible crop (e.g. potatoes, beets, maize, alfalfa, and beans) after wheat or barley crops; or other agronomic practices, such as 'thorough tillage'. However, if the soil has been managed with a monoculture of wheat, the disease is naturally reduced after a few years, a phenomenon known as 'Take-All Decline' (Lynch 1986). Antagonistic microorganisms are believed to build up against *G. graminis* var *tritici* under these conditions. An example of the effect of antagonistic microorganisms has been shown by inoculating wheat roots with a *P. fluorescens* strain, which produces the antibiotic phenazine-1-carboxylic acid, resulting in a reduction in severity of 'Take-All'

(Thomashow *et al.,* 1996). Antagonistic microorganisms, especially pseudomonads, are believed to build up in soils under cereal monoculture. This demonstrates the active role that pseudomonads play in the development of 'disease-suppressive soils' .

1.3.1.1 Biological control and nematode suppressive soils

There are numerous environmental biotic factors which can exert a pressure on the soil microbial community. Soil microorganisms exert a degree of control over populations of plant-parasitic nematodes. The cereal cyst nematode *Heterodera avenae* is a plant-parasite which causes commercial damage to crops in Northern Europe. Nematode-suppressive soils have been reported to reduce cereal cyst nematode populations by 90-95% (Kerry *et al.,* 1996); the fungus *Verticillium chlamydosporium* is known to be associated with these suppressive soils. *V. chlamydosporium* has been reported to be involved in the natural control of cyst nematodes in cereal monoculture in northern Europe and is important in the suppression of several other cyst nematode pests (Kerry *et al,* 1996). The establishment of a biological control agent is very complex, and because of the proven role of pseudomonads in other disease-suppressive soils, their role in effecting population shifts in nematode-suppressive soils merits further investigation. Future research, using the methodologies described in this thesis, should help to clarify our current understanding of microbial interactions in soils suppressive to plant pathogenic nematodes, as well as the plant-nematode-microorganism interactions.

1.3.1.2 Siderophores and suppressive soils

Bacteria have been reported to have an effect on the rhizosphere, via siderophore production. Siderophores are biologically-synthesised compounds which appear to be useful in reducing plant disease through antagonism of pathogens (Tate, 1995). The mode of action is effectively to complex the limited supply of iron (III) in soils, making it unavailable to pathogens. It has been shown that the presence of a *Pseudomonas* strain and/or its siderophore, pseudobactin, was correlated with a conducive soil becoming suppressive (Klopper *et al.,* 1980). This suggests that one factor in field suppressiveness to Take-All disease is microbial siderophores. However, siderophores are produced by a range of bacteria and fungi, so fungi may also have a role in these soils. Therefore, the positive effect of siderphore production by pseudomonad groups

must be assessed in conjunction with the rest of the soil microbial community, which will also contain pathogens, antagonists, PGPRs, siderophore and antibiotic producers.

1.3.1.3 Degradation of organic pollutants

Pseudomonads are amongst the most important organisms in the degradation of organic pollutants which occur within soils as a result of either intentional use (e.g. pesticide) or fallout of atmospheric pollutants (e.g. PAHs from combustion in power generation, or from vehicles). Any inhibition of these by heavy metals could reduce the rate of removal of organic pollutants in the soil, thus compromising an important soil function Rhizobia, which are important to the legumes that they nodulate, are known to be clearly affected by the heavy metals in sewage sludge (Giller *et al.,* 1998, Hirsch *et al.*, 1993, McGrath, 1987), yet the impact of contamination on a group considered generally more important to plants, like pseudomonads, has not yet been studied.

1.3.2 Pseudomonads: the keystone group?

Research on soil rhizosphere communities identifies one possible key group of related species, the fluorescent pseudomonad group, as they are environmentally important and have considerable economic impact. The development of methods for quantifying the diversity of soil pseudomonads will be required to allow for the complete investigation into this diverse group. Studying this fluorescent pseudomonad group, in the meantime, will provide valuable information into the effect of environmental stresses on soil micro-organisms and their balance in the bacterial community structure.

1.4 Environmental abiotic stress

Abiotic stress is caused by many environmental pollutants, but this study will concentrate only on the effect of heavy metal contamination in the soil microbial ecosystem. In the 1970s, many European countries introduced limits for heavy metal contamination in soils. Initial concerns concentrated on the control of heavy metal addition rates, via the disposal of contaminated sewage sludge on to agricultural soils. The application limits focused mainly on protecting against negative effects on crop plants, or animal grazing, and to protect the macro food chain from metal exposure (McGrath, 1987).

Sewage sludge contains higher concentrations of heavy metals than most soils, and regular application to land increases metal concentrations in the topsoil. This is because metals are not washed out of soils, and tend to accumulate in the surface layer. By 1997, about 50% of the total sludge produced annually in the UK was applied to agricultural land, and it is estimated that this will increase in future. By the end of 1998, dumping of sewage sludge by the UK in the North Sea ceased and much of this sludge must now be disposed of on to agricultural land. Most sewage sludges are contaminated with a cocktail of heavy metals, which may be deleterious to soil fertility and hence to agricultural productivity and subsequent food quality. For metal toxicity to occur, metals must be soluble in soil solution, and it is thought that the free metal ion is the most toxic chemical form to soil organisms.

It is important, therefore, to examine the effects of metal-contaminated sewage sludge applications to soil on microbial populations and their activities, as they contribute to long-term soil fertility.

Table 1.2 UK and EU limits for heavy metal contamination in PPM

Elements	UK limits	EU limits
Zinc	300	150-300
Copper	135	50-140
Nickel	75	30-75
Cadmium	3	1-3
Chromium	400*	100-150*
Lead	300	50-300

*(*These are currently under discussion)*

The effect of elevated heavy metal concentrations on soil micro-organisms was taken into consideration in legislation in the 1990s. The use of biologically-based observed effects on key sentinel species has guided UK policy, since the 1993 Independent Scientific Review (MAFF/DoE, 1993). The EU recommends a range within which heavy metal contamination may be tolerated, but individual countries are free to set

their own legal limits. The UK applies the upper limits permitted by the EU (Table 1.1; McGrath, 1994).

1.4.1 Heavy metal effects on microbial communities

The dynamics of heavy metal on soil properties is well documented. Properties such as pH, organic matter, clay content, iron content and Eh (redox potential) all alter the effects of metal contamination on soil microbes (Babich *et al.,* 1980). A long-term experiment, the **Market Garden Experiment** at Woburn Experimental Farm, Woburn, U.K., was sampled in order to minimise the variation of the residual effects of sewage sludge additions (i.e. equilibrated levels of contamination).

The soil is a sandy loam (Typic Udipsamment), pH 6.5 with 10% clay and 51% coarse sand. These plots received heavy metal contaminated sewage between 1942 and 1961 and their history is well documented. Alongside these are uncontaminated plots which received similar levels of organic matter inputs from farmyard manure (FYM) during the same period (McGrath *et al.,* 1987). The differences in soil properties were perceived to be minimal; on this basis, the observed differences between plots are thought, therefore, to reflect the influence of heavy metal contamination.

1.4.2 Chronic toxicity

The toxic effects of heavy metals on soil microorganisms are well reported (McGrath, 1994, 2001). Soil microorganisms differ in their sensitivity to metal toxicity, but when a critical level of exposure to heavy metals is reached, it will ultimately result in the death of cells. However, changes in microbial population due to increasing exposure to heavy metal, at subcritical levels, tends to be reflected by a gradual change in community structure.

The natural variability and inherent tolerance within microbial populations to metals are not obviously related to previous environmental exposure to heavy metals (Giller *et al.,* 1998). In previous studies, the Market Garden Experiment plots subjected to long term metal stress at modest levels did not maintain the same overall biomass, as in Farm Yard Manure (FYM) treated soils (Chander *et al.,* 1986). It was suggested that the biomass in the metal-contaminated soil had lower substrate utilisation efficiency as a

result of metal toxicity (Brooks and McGrath., 1994). Stressed soil microorganisms, diverting energy from growth to cell maintenance functions, thereby decreasing biomass (Killham 1985), may explain that phenomenon. But biomass is a measure of population density and is not designed to indicate changes within specific groups of organisms; to date, no clear relationship between the measure of diversity and biomass (process rate) has been shown (Bengtsson, 1998).

This raises a critical question: does the presence of heavy metal affect **diversity** of microbial groups in contaminated soils, or is the effect limited to microbial **numbers**? If the latter, then, rationally, one would expect key processes to remain; if the former, then important specific groups may be lost from the soil ecosystem. This happened in the Market Garden Experiment to rhizobia, which form nitrogen-fixing nodules in white clover (Hirsch *et al.,* 1993). The experiment is therefore selected for further investigation into the diversity and population dynamics of key groups, like pseudomonads, in abiotic stressed soils.

1.5 Diversity and population dynamics

Research on the effect of environmental stress (both abiotic and biotic) on soil has concentrated on crop production and crop yields. This study will investigate in depth the effect of an abiotic stress (e.g. environmental pollutants) on the soil's bacterial community structure. Many complex factors are thought to account for changes in the bacterial dynamics and community structure. A review of relevant research indicates a possible group of key species, fluorescent pseudomonads, which will be investigated in order to gain greater understanding of the complex ecological interaction within bacterial communities. Assessment of genetic diversity, and the genetic relatedness of strains within natural populations, will yield valuable information about bacterial population structure and how groups adapt to environmental stresses (Niemann *et al.*, 1997).

The detrimental effect of high concentrations of heavy metal is widely accepted, but the effect of gradual metal toxicity on soil microbial communities is unknown. Mechanisms of metal toxicity and tolerance in microorganisms at molecular level are unclear (Giller *et al.,* 1998). A major perturbation in the rhizobial population was seen in heavy metal

contaminated soils from the Woburn Market Garden Experiment (Hirsch *et al.,* 1996). If other populations were also effected, it would imply serious consequences for the growth of crop plants on metal-stressed soil, although none were found in the Market Garden Experiment, with the exception of white clover.

Other studies have shown the effect of pesticides on pseudomonad populations in soils. These pesticide-treated soils showed no significant effect on the total **number** of pseudomonad groups, but a **genetic** shift in the pseudomonad population was shown, indicating either a direct or indirect effect of pesticides on the distribution of bacterial strains (Nicholson *et al.,* 1997). However, no adverse (or beneficial) effects to plants were associated with these differences.

These findings are important, because they imply that the effect of stresses on the microbial community may be either as drastic as the effect observed on *Rhizobium* populations (Hirsch *et al.,* 1996), or as subtle as the effect on the pseudomonad population (Nicholson *et al.,* 1997). It is imperative, therefore, that both numbers and diversity of a group are assessed, along with functional measurements, to establish a fuller understanding of how severely stress affects microbial populations soil.

1.5.1 Model of stressed ecosystems

Models are useful in examining soil population dynamics; one proposed by Austin and Smith describing plant populations in stressed environments can be used for studying soil populations (Austin and Smith, 1989). Their model of stress relationship and population shift (elaborated further by Giller *et al.,* 1998), was extrapolated from the plant kingdom. It proposed two hypotheses: if the microbial community responded in a similar manner to increasing environmental stresses, the response could be either:

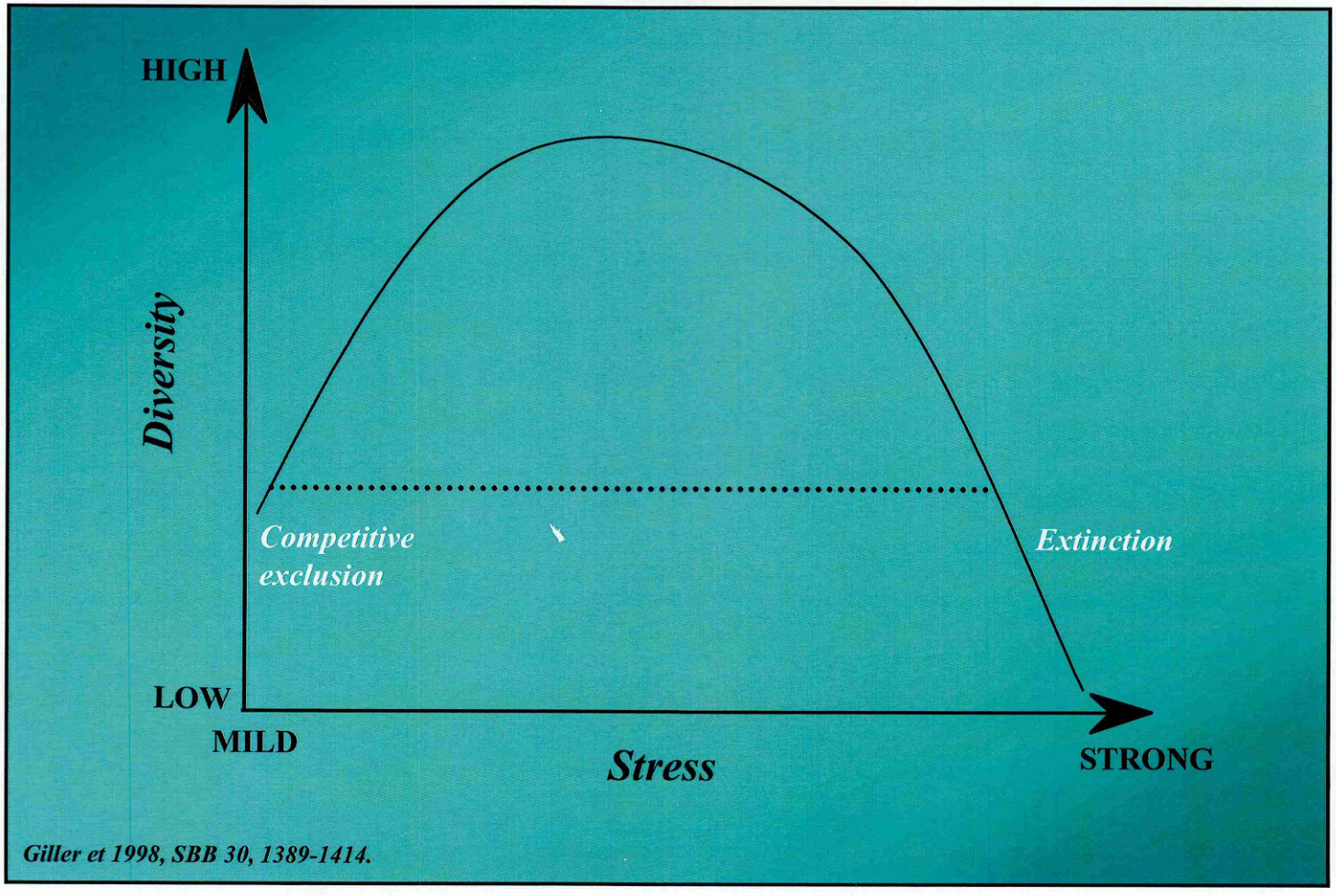
- a linear response to stress, i.e. a straight line decrease, in response to an increase in stress;
- or a bell shape response (i.e. as stress increases, it yields a pattern of low level diversity, followed by an increase in diversity, finally followed by a crash in diversity; Figure 1.6). The theory behind such a response is that when organisms are exposed to stress, they may be freed from a process known as *competitive exclusion* (which otherwise permits only selective species dominance; Lynch and Hobbie,

1988), thereby allowing more species to occupy an environmental niche, before an accumulation of stress precipitates a permanent and catastrophic decline in numbers.

This model suggests that in a healthy plant ecosystem, there are a few dominant groups of organisms sitting on top of a species-rich underlying population. As the system becomes stressed, dominant organisms lose their adapted advantage and the apparent number of plant species increases, indicating changes in the system, leading to a possible breakdown. When the level of stress becomes critical, the ecosystem structure crashes, leaving only a few dominant groups able to cope with extreme environmental conditions. There is no longer a diverse underlying population to buffer the system. With the previously rich diversity within the ecosystem reduced there is only a basal background with a much lower overall diversity. According to this model, the system may *initially* seem to survive and function normally, but problems occur if the system is then exposed to a second (and conceivably lower level) stress, which may even be a simple water or nutrient stress.

If a natural reservoir of microflora is lost (Giller, *et al.*, 1998) a system may have difficulty compensating for stresses. Where there is moderate stress (e.g. from heavy metals), organisms may be freed to some extent from a process known as competitive exclusion; the presence of little stress, in contrast, allows only few 'bacterial types' to dominate and, **unless the underlying diversity is measured,** it appears that there is limited 'species richness'. (This highlights the importance of assessing both functional *and* genetic diversity in stressed environments). This effect is weakened by stress, allowing more types to be observed as species become more 'even' (Giller *et al.,* 1998), as there is a breakdown in the natural control of the ecosystem. The increase in

Figure 1.6 The hump-backed relationship between species diversity and disturbance



observed species can lead to a misleading assumption that 'species richness' has increased. In fact, that can be a simple function of the way bacterial groups are studied.

This model was used to describe the distribution of *Rhizobium* plasmid profiles in contaminated soils (Turner *et al.,* 1996), but care must be applied as the model's assumptions are based on species diversity, not plasmid diversity. However, there is a strong possibility that the model predication may be effectively applied to other components of the micro-ecosystem and may reflect population dynamics in pseudomonad groups under abiotic stresses. The methodology developed for this thesis should make it possible to investigate potential changes in diversity of soil microbial communities in stressed soils. The use of a defined group (i.e. fluorescent pseudomonads) will allow assessment of changes in diversity caused by know levels of heavy metals in stressed soils. Results can then be compared to those with a known 'indicator' (ie Rhizobium) to demonstrate whether the pseudomonads are in fact a potential indicator for stress.

1.6 Summary

The Report of the Nineteenth Royal Commission recommended that a Soil Protection Policy be drawn up and implemented for the UK. This means that acceptable limits for soil metal contamination, which do not cause adverse effects on the soil ecosystem, need to be decided. In many cases, suitable assessment techniques do not exist, yet a national soil quality monitoring scheme requires the establishment of meaningful chemical, physical and biological techniques for assessing soil quality.

This project intends to addresses the biological effects of heavy metals on soil microbes, which are important for the fertility and sustainability of agricultural systems. This research will address the development of "biological measures" of soil quality, which can be used to shape Soil Protection Policy, including (eventually) the determination of acceptable levels of sewage sludge disposal on land, in addition to a programme of monitoring and surveillance.

As indicated at the start of this chapter, our knowledge both of soil microbial populations (which are vast in number *and* hugely diverse), and the precise implications of a loss of diversity, are currently constrained. Accordingly, a deeper knowledge of soil microbial diversity needs to be established before I can reliably determine if diversity can be used as an appropriate indicator of soil quality. In the meantime, our provisional view is intentionally conservative: that because diversity, both functional and genetic, is important in preserving ecosystems and agriculture, soil microbial diversity should be preserved in its present condition.

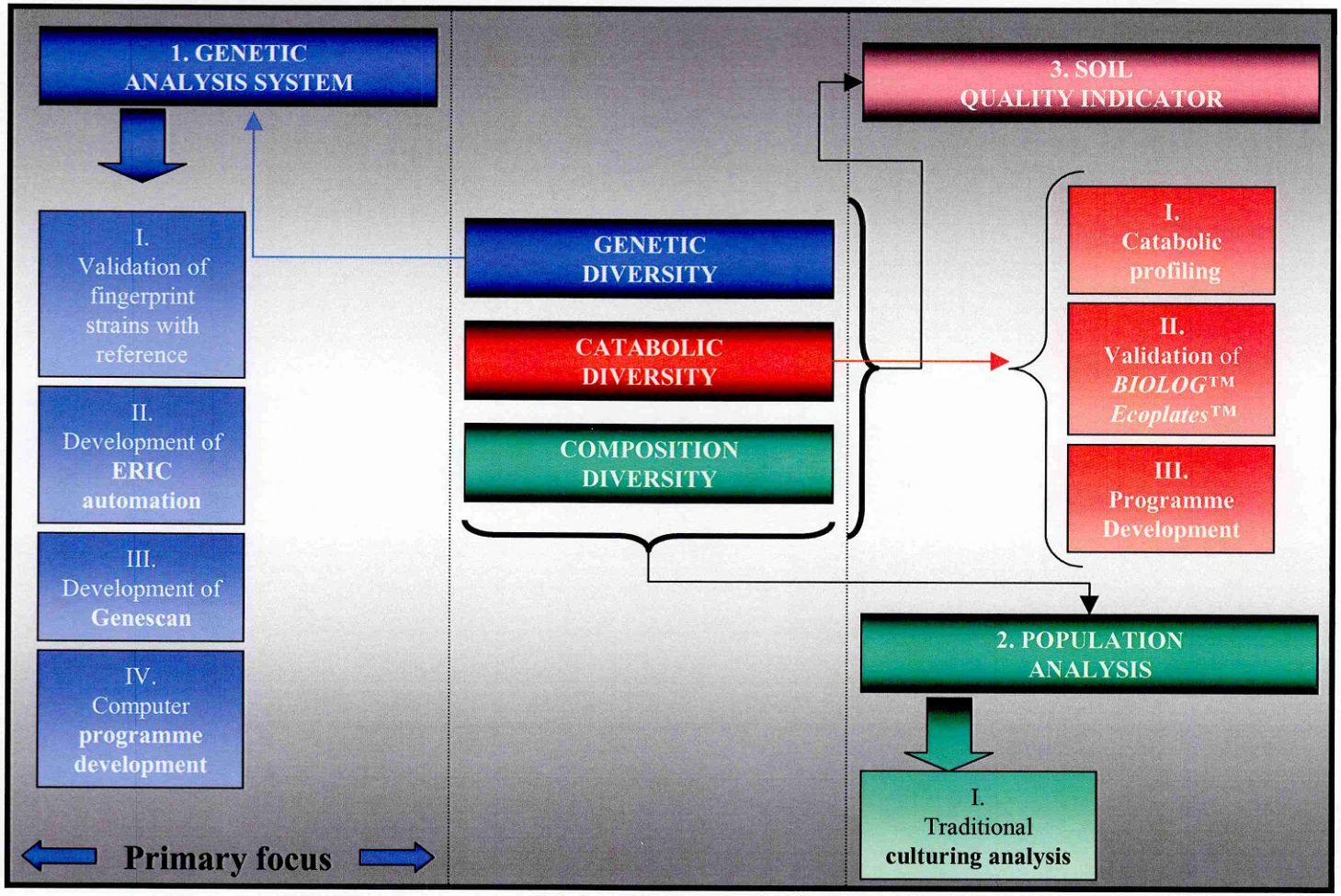
Past research has concentrated on the effects of environmental stresses on the soil, as measured by crop production and yields. In this project, however, I question the effect of an abiotic stress -i.e. environmental pollution- on the structure of the soil's bacterial community. Because many complex factors influence changes in bacterial community structure, this study has chosen to concentrate on the ecology of one of the most important bacterial groups, the pseudomonad group. The assessment of the genetic diversity and relatedness of strains within natural populations will provide information about bacterial population structure and how it is subsequently affected by environmental stress.

1.7 Project objectives

The objectives of this research, illustrated in the flow diagram below (Figure 1.7), were threefold:

1. **To develop a system capable of studying genetic diversity in the fluorescent pseudomonad group of soil bacteria.**
2. **To assess if the three different measurements used to profile microbial diversity** (catabolic; genetic; population size and composition) **could be used to study population shifts in soil.**
3. **To assess if a combination of these measurements could be used as an indicator of soil quality.**

Fig 1.7. Research modules and objectives



Chapter Two:
General Methods

2.1 Introduction

This chapter describes routine methods. Methods specific to other sections in this research, or which have been modified extensively, will be dealt with when mentioned in the relevant chapter and are not examined here.

2.2 Site and soil sampling

2.2.1 Site description

The Woburn Experimental Farm, Woburn, U.K, is the site of the Market Garden Experiment (Figure 2.1). The soil is a sandy loam (typic Udipsamment), pH 6.5, with 10% clay and 51% coarse sand. Sewage sludge was applied between 1942 and 1961 and the microbiological properties of the sludge-amended soil are compared with that of a soil which had been amended with farmyard Manure (FYM) between 1942 and 1965. The plots have been set to grass since 1989.

These, then, are the FYM and sewage plots referred to throughout this thesis. Both plots are categorised as **Series A, number 2 (FYM)** and **number 3 (sewage sludge)**. The two plots sampled from Series A were:

1. **FYM plot 2** (94/w/rn/4 series A_02) which received 10.4 tonnes organic matter ha^{-1} yr^{-1} between 1942 and 1961.
2. **Sewage sludge plot 3** (94/w/rn/4 series A_03) which received 16.4 tonnes organic matter ha^{-1} yr^{-1} between 1942 and 1961.

2.2.2 Sampling regime

Soil samples were taken with a 1 cm-soil corer. 10 sub-samples were taken and all cores were mixed and sieved with a 2 mm sieve. Multiple soil samples were taken and mixed together to account for the spatial variation that occurs within the soil environment. These soil samples were used in the following procedures. Soil samples were processed and tested on the same day for moisture content by overnight oven drying at 80 ° C.

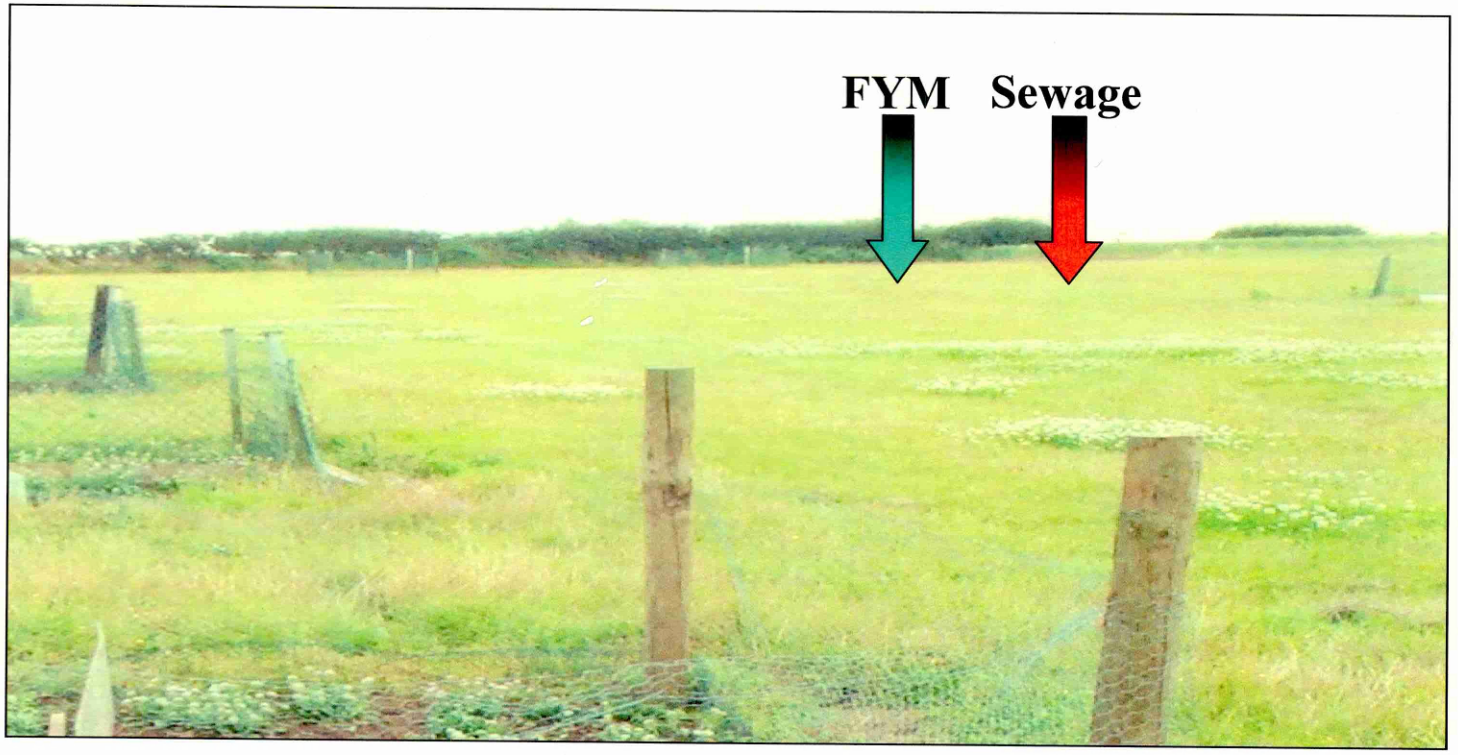


Fig 2.1. A ground level view of the Market Garden Experiment, showing both FYM and Sewage plots

Farm Yard Manure (FYM): 10.4 tonnes organic matter ha^{-1} yr^{-1}.
Sewage Sludge (Sewage): 16.4 tonnes of heavy metal contaminated organic matter ha^{-1} yr^{-1} between 1942 and 1961.

Soils were stored at 4 °C between mixing and inoculation of BIOLOG GN1™ plates. However, storage time was kept to a minimum, i.e. not more than 4 days.

2.3 Soil chemistry

The total metal content of the soils and the ammonium nitrate extractable metals were measured according to the method of McGrath and Cunliffe (1985).

2.3.1 Soil pH

Soil solutions were prepared with 1 g of sieved soil in 9 ml of sterile distilled water and the soil solution pH was measured using an Orion 720A pH meter (Orion, Cambridge, UK). All assays were done in triplicate.

2.3.2 Preparation of soil samples

Fresh soil was dried in a cabinet with a fan circulating air at room temperature. The air-dried soils were milled in an agate mill as the soils were for metal analysis. Care was taken to avoid contaminating samples and the milled soils were packed in 0.5 kg bags, labelled and stored in boxes.

2.3.3 Aqua regia acid digestion

Digestion of soils with aqua regia acid is used for the determination of major and trace elements. The samples are digested in glass test-tubes which are heated in time/temperature controlled heating blocks, filtered and then analysed by Inductively Coupled Plasma Emission Spectrometer (ICP). This method does not give absolute total contents; it does give results sufficiently close to accepted values for different types of soils and sludges for it to be of value in the routine monitoring of metal contents. If Cadmium levels fell below the accuracy specification of the ICP, then the graphite furnace analysis was used.

Procedure

250 mg of finely milled air-dried soil was put into a 25ml graduated digestion tube. The following additions were made: 4 ml hydrochloric acid (AR s.g. 1.18) and 1 ml nitric acid (AR s.g. 1.42). This equates to 5 ml aqua regia. All samples were left to stand at room temperature for 2 hours (blanks and standard soils were included). Then samples were placed in the heating block and were digested using the following heating programme:

Ramp No.	Temp. Rise sec °C-1	Dwell Time hours	Dwell Temp. ° C
1	20	2	25
2	20	3	60
3	20	1	105
4	20	2	125

The samples were mixed occasionally at or below 60 °C to prevent a cap of partly digested material rising up the tube and then the above heating programme was carried out overnight. The tubes were removed from the heating block and 5 ml of 25% hydrochloric acid was added to each tube, whirlimixed and reheated to 80 °C for 30 min. They were then removed from the block and allowed to cool. Once cooled, the samples were made up to 25 ml with deionised water. The samples were whirlimixed and filtered through a Whatman no. 40 filter paper in to a Sterilin vial (the first 5 ml of filtrate was discarded).

2.3.4 Inductively coupled plasma emission spectrometer (ICP-MS)

Inductively coupled plasma emission spectrometer (ICP-MS) is the combination of inductively coupled plasma with a mass spectrometer. The ICP uses argon to generate efficiently singly-charged ions from the elemental species within a sample (samples are pre-treated in the aqua regia to allow heavy metal content to be estimated). These ions are then directed into a quadrupole mass spectrometer, which permits the quantification

of elemental isotopic concentrations and ratios. The ICP-MS consists of a sample introduction system (nebulizer, spray chamber and pump), the plasma and spectrometer. The sample introduction system brings the sample solution in the form of an aerosol, to the plasma. The plasma is a highly ionised, hot gas, which is stable, and chemically inert with temperatures near 9,725 °C. Light from the plasma is accepted by the spectrometer and separated into components and detected by photomultiplier tubes; the individual components are then recorded. Metal concentrations were determined by *inductively coupled plasma atomic emission spectrometry* (ICP-AES, Accuris, Fisons Instruments, East Grinsted, UK).

2.3.5 Graphite furnace atomic absorption

Because Cadmium levels fell to the lower limit of the accuracy specification of the ICP-AES, the samples were analysed on the graphite furnace to ensure that the correct levels of Cadmium were being recorded.

Graphite Furnace Atomic Absorption (**GFAA**), has an electrically heated graphite tube. The samples are introduced directly into it. The entire sample is atomised, and the atoms are retained within the tube for an extended period. As a result, sensitivity and detection limits are significantly improved in the GFAA, but analysis times are longer, and fewer elements can be determined, compared with older graphite flame furnace methods. However, this enhanced sensitivity and ability to analyse very small samples significantly expands the capabilities of atomic absorption. Its advantage is demonstrable with low levels of Cadmium, as it is a reliable method for Cadmium levels at 5 ppm or lower.

2.3.6 LECO combustion analyser operating procedure

The LECO FP2000 Combustion Analyser (Leco, Stockport UK) is a fully automatic instrument used for the determination of total nitrogen and carbon in soils and is a modified version of 'Dumas' digestion method. Soil Samples were treated as described in section 2.3.1. Weighted samples are placed in the auto-loader and pushed into the combustion chamber where the furnace and flow of oxygen gas cause the sample to combust. The combustion process converts any elemental C, S and N to CO_2, SO_2, N_2

and NO_x. These gasses then pass through two anhydrone tubes (to remove H_2O), a particle filter, and collect in a ballast tank. The gas is left to equilibrate before being released into an aliquot loop and through the infrared cells where carbon and sulphur are detected. Gas passes from the aliquot loop to the catalyst heater where NO_x is reduced to N_2, then through Lecosorb to remove CO_2 and anhydrone to remove H_2O. The remaining N_2 and helium carrier gas flow through a thermo conductivity cell where the nitrogen is measured. The % N and C are then assessed according to the manufacturer's protocols.

2.3.7 Carbonate in soils

Soil is treated with hydrochloric acid, which liberates the carbonate in the form of carbon dioxide. The amount of carbonate in the soil is calculated by comparing the pressure produced by the sample against the pressures produced by known weights of calcium carbonate. The resulting carbon may then be subtracted from the total carbon figure obtained from the Leco combustion analyser to give an organic carbon value.

Procedure

5g of finely ground soil (section 2.3.1) was weighed into a polythene cup and 4 ml of boiled distilled water (boiled to expel dissolved CO_2 and in a CO_2- atmosphere) was added. The polythene cup containing the soil was placed inside the reaction bottle which contained 25 ml 2M HCl and 1 ml distilled H_2O. The bottle was sealed, placed on the flask shaker and shaken thoroughly for 3 minutes to mix soil and acid. A pressure reading was taken by slowly opening the manometer and the pressure reading was recorded. The manometer was calibrated via a range of pure $CaCO_3$ AR (concentration 0.1g to 0.5g) which was plotted to give a standard curve of pressure.

The amount of $CaCO_3$ in the soil was calculated as follows:

$$\frac{\text{Amount } CaCO_3 \text{ in soil} \times 100}{\text{Weight of soil sample}} = \% \ CaCO_3 \text{ in sample}$$

2.4 Carbon utilisation

2.4.1 BIOLOG™

The BIOLOG GN1™ Gram Negative (**GN**), micro plates were developed as a microbial identification system and can identify more than 569 species including enteric, nonfementer and fastidious strains (Bochner, 1989). However, the environmental application of this system uses them to study biological diversity in soil (Garland,1996; O'Flaherty *et al.*,1998), via the carbon utilisation profiles, comparing the different catabolic potential of different microbial communities. The BIOLOG GN1™ plates contain 96 wells of which one is a reference well and 95 are test wells containing carbon substrates (Table 2.1). Every well also contains a complex, low concentration, nutrient medium along with a Redox dye. The Redox dye chemistry works on the principle that regardless of its molecular structure, virtually any chemical substrate that is oxidised by a cell, will result in the formation of NADH, which donates electrons to the electron transport chain. Redox dyes such as tetrazolium tap electrons from this flow, converting it to a highly coloured purple formation (Bochner, 1989).

Table 2.1 List of substrates for the 96 well BIOLOG™

1	blank	33	turanose	65	D-alanine
2	α-Cylodextrin	34	xylitol	66	L-alanine
3	dextrin	35	methyl pyruvate	67	L-alanyl-glycine
4	glycogen	36	mono-methyl succinate	68	L-asparagine
5	tween 40	37	acetic acid	69	L-aspartic acid
6	tween 80	38	cis-aconitic acis	70	L-glutamic acid
7	N-acetyl-D-galactosamine	39	citric acid	71	glycyl-L-aspartic acid
8	N-acetyl-D-glucosamine	40	formic acid	72	glycyl-L-glutamic acid
9	adonitol	41	D-galacturonic acid lactone	73	L-histidine
10	L-arabinose	42	D-galacturonic acid	74	hydroxy L-proline
11	D-arabitol	43	D-gluconic acid	75	L-leucine
12	cellobiose	44	D-glucosaminic acid	76	L-ornithine
13	i-erythritol	45	Dglucuronic acid	77	L-phenylalanine
14	D-fructose	46	α-hydroxybutyric acid	78	L-proline
15	L-fucose	47	β-hydroxybutyric acid	79	L-pyroglutamic acid
16	D-galactose	48	γ-hydroxybutyric acid	80	D-serine
17	gentiobiose	49	p-hydroxy phenylacetic acid	81	L-serine
18	α-D-glucose	50	itaconic acid	82	L-threonine
19	m-inositol	51	α-keto butyric acid	83	D,L-carnitine
20	α-D-lactose	52	α-keto glutaric acid	84	γ-amino butyric acid
21	lactulose	53	α-keto valeric acid	85	urocanic acid
22	maltose	54	D,L-lactic acid	86	inosine

23 D-mannitol	55	malonic acid	87	uridine
24 D-mannose	56	propionic acid	88	thymidine
25 D-mellibiose	57	quinic acid	89	phenyl ethylamine
26 β-methyl-D-glucoside	58	D-saccharic acid	90	putrescine
27 Dpsicose	59	sebacic	91	2-amino ethanol
28 D-raffinose	60	succinic acid	92	2,3- butanediol
29 L-rhamnose	61	bromo succinic acid	93	glycerol
30 D-sorbitol	62	succinamic acid	94	D,L-a-glycerol phosphate
31 sucrose	63	glucuronamide	95	glucose-1-phosphate
32 D-trehalose	64	alaninamide	96	glucose-6-phosphate

Utilisation of a carbon source is detected as an increase in the respiration of living cells in the well, leading to an irreversible reduction of a tetrazolium dye. A positive utilisation reaction is indicated by purple colour formation in a well (Bochner, 1989) but if the cell cannot oxidise a carbon source, its metabolic rate does not increase and no colour changes occur. The test yields a pattern of purple wells, which constitutes a 'metabolic fingerprint' of the capacities of the inoculated organism (or of the microbial community).

2.4.2 EcoPlates™

A new generation of BIOLG GN1™ plates were released named EcoMicroPlates ™, although they work on the same principle as BIOLOG GN1™ plates. These EcoPlates™ were designed to have "ecologically relevant" substrates: they contain one reference well and 31 sole carbon sources (Table 2.2); the test has three replicates in each 96 well. Each well contains the redox dye tetrazolium, which is reduced by NADH produced by respiration pathways. The EcoPlates™ were used in a previous studies as a fast screening technique to detect changes in populations attributed to stress (Knight *et al.*, 1997).

Table 2.2 Substrates on the 32 well EcoMicroPlates™

1 Water	12 L-Phenylalanine	23 Itaconic Acid
2 β-Methyl-D-Glucoside	13 Tween 80	24 Gylcyl-L-glutamic-acid
3 D-Galactonic Acid γ-Lactone	14 D-Mannitol	25 D-cellobiose
4 L-Arginine	15 4-Hydroxy Benzoic Acid	26 Glucose-1-Phosphate
5 Pyruvic Acid Methyl Ester	16 L-Serine	27 α-keto Butyric acid
6 D-xylose	17 α-Cyclodextrin	28 Phenylethyl-amine
7 D-Galacturonic Acid	18 N-Acetyl-D-Glucosamine	29 α-D-Lactose
8 L-Asparagine	19 o-Hydroxybutyric acid	30 D,L-α-Glycerol Phosphate
9 Tween 40	20 L-Threonine	31 D-Malic Acid
10 I-Erythritol	21 Glycogen	32 Putrescine
11 2-hydroxy Benzoic Acid	22 D-Glucosaminic Acid	

2.4.3 BIOLOG GN1™ and EcoPlates™ method

BIOLOG GN1™ and EcoPlates™ plates were inoculated using a standard method (Garland and Mills 1991) and each plate was inoculated with a diluted soil extract (section 2.5.2). Bacterial numbers in the soil were established via a viable count on the sampling day and soil samples were stored at 4° C until numbers were estimated. Then fresh soil suspensions were serially diluted to give approximately 10^4 cfu per final cell concentration count in the BIOLOG GN1™ assay. Soil samples were normalised for the number of bacteria using information from a preliminary count on TSA before inoculation, avoiding the need to correct results by dividing by the average well colour development (AWCD) (Garland 1987). The plates were inoculated with 145 μl aliquots of the soil suspension in each well with three replicates to allow for statistical analysis. The plates were wrapped in cling film to avoid moisture loss, and incubated at 28 °C for 7 days (note that the plates were equilibrated at room temperature for 60 min before use).

Once inoculated, the plates were then read at 590 nm immediately (0 hours) on the Labsystems Multiskan RC reader (Labsystems, Basingstoke, UK) RC version 4.0 plate reader and every 12 hours approximately until colour formation was complete (or 7 days incubation). The data recorded was used to assess any functional changes in microbial communities via the soil's carbon substrate utilisation ability. To aid analysis, two Genstat programs were written to analyse the BIOLOG GN1™ and EcoPlates™ (Appendix A; diversity indices and statistical approaches are described in Chapter 4.)

2.5 Microbiology methods

2.5.1 Quantification of bacteria

Soil samples are taken as described in section 2.2.2. The microbial population analyses were performed on the day of sampling (where possible) to allow an accurate picture of the distribution of the cultural soil micro-organism in the soil and dry weights were performed. Soil for the BIOLOG GN1™ test was stored at 4 °C to allow for the normalised technique; storage time was kept to a minimum (48 hours). Soil samples were stored at -80 °C so that further DNA analysis could be performed if necessary at a later date, or as required for further projects.

2.5.2 Preparation of soil extracts

One gram of pre-treated soil sample was taken and resuspended in 10 ml of sterile distilled water (dH_2O). The soil samples were vortexed for 7 min to extract soil microrganisms from soil particles (Standard Laboratory Practice at IACR-Rothamsted).

2.5.2.1 Protocol for serial dilution (viable count)

For serial dilutions 1 ml of suspended soil sample was added to 9 ml sterile dH_20 to make a one in 10 dilution (10^{-1}). One ml of this dilution was added to a further 9 ml solution to make a one in 100 dilution (10^{-2}). This was repeated until 10^{-7} dilution was reached. Triplicate samples of 100 µl of each dilution were spread onto agar plates for microbial counts using standard procedures (Alef & Nanninpieri, 1996), and 20 µl for drop plates (Miles and Mirsa, 1949). The diluted soil extract was plated out on the range of bacterial media and PDA agar plates for fungi, and incubated at 28 °C. Colonies were counted after 24 hr and 48 hrs incubation for bacterial cultures, and fungal counts monitored between 2 and 5 days with counts recorded at 48 hrs.

CFU ml^{-1} was calculated using the formula below:

$$\frac{\textbf{Mean count}}{\textbf{(Dilution used x amount plated out \{ml\})}}$$

2.5.3 Media preparation

To culture soil micro-organism from soil a selection of commercial selective media are used as follows:

- **Tryptone Soy Broth Agar:** (TSA, Oxoid, UK) and 1/10th strength TSA selected for heterotrophic bacteria
- **Nutrient Agar:** (NA, Oxoid, UK) pre-treated at 50 °C for 2h before incubation for heat resistant spore-forming bacteria
- **Pseudomonad Selective Agar:** (PSA, Oxoid, UK) used to estimate the number of fluorescent pseudomonads in the total culturable bacterial heterotrophic population. It contains the CFC antibiotic supplement at a final concentration in the agar of 10 $\mu g\ ml^{-1}$ centrimide, 10 $\mu g\ ml^{-1}$ fucidin and 50 $\mu g\ ml^{-1}$ cephaloridine
- **MacConkey Agar:** (MCA, Oxoid, UK) selects putative enterobacteria (found typically in manure contaminated by human or animal manure).

These media were chosen for their ability to select a variety of putative microbial sub-populations and were used to study the size and composition of the soils populations. This permitted assessment of the effect of heavy metals on, and composition of, the culturable components of microbial communities.

2.6 Molecular biology

2.6.1 Extraction of bacterial DNA

Total bacterial genomic DNA was extracted from pure cultures in accordance with a protocol described by Sambrook (Sambrook *et al.,* 1989; Chen *et al.,* 1993) and a modified plant DNA extraction technique (Murray and Thompson 1980). Bacteria from a culture grown in liquid media till mid-log phase were lysed using 10 % solution of dodecyl sulphate (SDS) in TE buffer. Proteins were removed by enzymatic digestion with 100 $\mu g\ ml^{-1}$ proteinase K (Sigma) for 1h at 37 °C. Cell wall debris, polysaccharides, and remaining proteins were removed by selective precipitation with CTAB in 0.7 M NaCl buffer. CTAB–protein complexes were removed by chloroform/isoamyl alcohol (v/v 24:1) extraction and the nucleic acid was cleaned by extraction with equal volumes of phenol/chloroform/isoamyl alcohol. Nucleic acids

were precipitated by the addition of 0.6 vol. isoproponal, incubated on ice for 20 min, then centrifuged at 9000 rpm for 10 min at 4 °C. The pellet was washed with 1 vol. of 70 % ice cold ethanol to remove traces of isopropanol, air dried for 10 min before re-suspension in an appropriate volume of 10 mM Tris-HCl (pH 8) buffer.

2.6.2 Fast bacterial DNA extraction

The bacterial cultures were grown up on a selected media to 3 to 4 days old. Then 5-10 bacterial colonies were taken, with a sterile tooth pick and carefully resuspended in 50 µl of lysis buffer containing 0.05 M NaOH and 0.25% SDS (the lysis buffer must be preheated to 37 ° C before use to ensure that the SDS is in solution). The solutions were then vortexed, heated for 15 min at 95°C in a thermal cycler, and cooled to 10°C. This releases the DNA from the cells into the solution. The samples were then centrifuged for 5 min at 13000 rpm to remove cell debris and then 10 µl of the clear lysed solution was diluted (1/10 in dH_2O molecular biology grade water). In the standard PCR reactions, 1 µl of this dilute solution was used as the DNA sample. The DNA was quantified by electrophoresis and then stored at –20 °C until required.

2.6.3 Gel electrophoresis

2.6.3.1 Agarose gels

The standard method used to separate, identify, and purify DNA fragments is electrophoresis through agarose gels (Ultra Pure DNA Grade Agarose, Bio-Rad). The percentage of the agarose ranged between 0.8 to 3 % (depending on fragment sizes). The samples were mixed with a loading buffer at 10:1 (0.25% bromophenol blue, 0.25 % xylene cyanol and 30 % glycerol in H_2O), and run in Tris-Borate buffer. (**TBE**) (Tris 0. 089 M, 0.089 M boric acid, 0.002 M EDTA at pH 8.0).

2.6.3.2 Visualising DNA

The most convenient method of visualising DNA in agarose gels is the use of the fluorescent dye ethidium bromide (Sharp 1973, Sambrook *et al.*, 1989). Ethidium bromide contains a planar group that intercalates between the stacked bases of DNA. This causes dye bound to DNA to display an increased fluorescence compared to dye in free solution. Ultraviolet (UV) radiation is absorbed by the DNA at 260 nm and

transmitted to the dye, which fluoresces at 320nm, resulting in a visible red colour. Ethidium bromide can be used to detect both single and double stranded nucleic acids (both DNA and RNA). Within certain limits, the mobility of the DNA fragment is determined by its size; the larger the fragment, the slower it moves through the gel. The higher the concentration of the gel, the greater it impedes the movement of the DNA. Therefore, gels (2.5-3%) were used for visualising only small DNA products, such as PCR products, whereas less concentrated gels (0.8-1.5 %) were used to visualise larger fragments, such as DNA preparations and total genomic digests. As little as 10 ng can be detected on gels through fluorescence after staining, which can be compared to a series of standards to estimate the concentration of a preparation.

2.6.3.3 Staining DNA in agarose gels

The ethidium bromide may be incorporated into the gel. This affects the mobility of linear duplex DNA by approximately 15 per cent. However ethidium bromide-free gels can be run with ethidium bromide staining after electrophoresis is completed, which is the preferred method when examining PCR fingerprinting profiles. The gel is immersed in dH_2O containing ethidium bromide (0.5 $\mu g\ ml^{-1}$ in dH_2O) for approximately 15 minutes at room temperature and then destained in dH_2O at room temperature for approximately 30 minutes and photographed over a 320 nm UV transilluminator with Polaroid type 667 positive/negative film.

2.7 Polymerase chain reaction (PCR)

PCR was used to amplify a segment of DNA which lies between two regions of known sequences. Oligonucleotides were designed to act as primers and bind to the homologous regions. The oligonucleotides typically had different sequences and were complementary to sequences that a) lie on opposite strands of the template DNA and b) flank the segment of DNA to be amplified. The template DNA was first denatured by heating in the presence of a large excess of primers and the four dNTPs (dATP, dCTP, dGTP, dTTP). The mixture was cooled and the primers annealed to their target sequences, after which the annealed primers were extended with DNA polymerase at a temperature optimum for the enzyme to function. The cycle of denaturing, annealing and extension was then repeated many times. Because the products of one round of

amplification serve as templates for the next, there was an exponential increase in the level of target DNA. Therefore, small quantities of DNA can be amplified and detected using this method

2.7.1 Protocol for PCR

The standard conditions for PCR are noted here, specific annealing temperatures or modifications are referred to when mentioned in the text. PCRs were performed in 20 µl total volume. Each solution contained 1 µl template DNA (20–60 ng), 2 µl x 10 PCR reaction buffer, containing Mg^{2+} at conc. of 1.5 mM (Boehringer Mannheim), 2 µl of each primer (0.02 pM), 200 µM of each dNTP and 0.4 U of Taq (Boehringer Mannheim). Mineral oil (Sigma) was added to prevent evaporation. DNA amplification was performed in a thermal cycler (TRIO-Thermoblock TBI, Biometra). The general parameters for PCR were 30 cycles of DNA denaturing at 94 °C for 1 min, primer annealing (temperature specific for each reaction) for 1 min, and extension at 72 °C for 1 min. A final extension of 72 °C for 5 min was added to each reaction before it was cooled to 4 °C and maintained. Control reactions were prepared in the same way but without template DNA. PCR products were observed by electrophoresis on a 2.5 % agarose gel.

2.8 16 S Ribosomal (rRNA) PCR

The ribosomes (comprising RNA and proteins) are a key element of the protein-synthesising machinery in cells and are vital to all organisms. They consist of three small subunit rRNAs: 5S rRNA, 16S rRNA and 23S rRNA (section 1.2.3.2) together with ribosomal proteins. The gene sequences (rDNA) of the 16S rRNA subunits are conserved and are believed to demonstrate phylogenetic relationships between species (Woese, 1987). The 16S rRNA PCR methods are of interest because they are simple, rapid and universally applicable. In this study two sets of oligonucleotides primers are used: the 16S Universal Primers and the 16S Pseudomonad Specific Primers.

The PCR products are taken and digested with a range of restriction enzymes and the amplified-rDNA restriction analysis generates mostly species-specific patterns (reflecting the conserved nature of the rRNA genes). This method of analysis is called

ARDRA (amplified ribosomal DNA-restriction analysis, [Vaneechoutte *et al.*, 1993]), which has the same principle as RFLP (restriction fragment length polymorphism) analysis, except that a PCR-mediated amplification of particular DNA fragments precedes restriction analysis. The ARDRA allows a simple classification of bacterial strains on the restriction digest profiles of the 16S rRNA PCR product.

2.8.1 16S Universal primers

Primers	Sequence (5'-3')	Designed for
fD1	**AGA GTT TGA TCC TGG CTC AG**	Most eubacteria
rD1	**AAG GAG GTG ATC CAG CC**	Most eubacteria

The standard PCR protocol was employed (section 2.7.1) with the following modifications: the DNA concentration was set at approx. 50 μg, dNTPs was 200 μM, Primers fd1 and rD1 at 50 pM. The Taq (Boehringer Mannheim) was 2.5 units per reaction in 25 ml volume. This method has been modified from the method outlined by Weisburg *et al.,* (1991).

2.8.1.1 PCR program

The 16S Universal parameters for PCR were: DNA denaturing at 93 °C for 3 min, then 35 cycles of DNA denaturing at 94 °C for 1 min, primer annealing at 55 °C for 1 min, and extension at 72 °C for 2 min. There was then a final extension of 72 °C for 5 min before being cooled to 4 °C; samples were then stored at –20° C till required.

2.8.2 16S Pseudomonad specific primers

16S Pseudomonad Specific Primers are defined by Widmer *et al.,* (1998) are as follows:

Primers	5'-3'	
Ps- for	**GGT CTG AGA GGA TGA TCA GT**	20-MER
PS- rev	**TTA GCT CCA CCT CGC GGC**	18-MER

(where Ps- forwards; PS- reverse)

The 16S Pseudomonad Specific PCR protocol was employed with the following modifications: the DNA concentration was set at approx. 50 μg, dNTPs was 200 μM and primers Ps-for and PS-rev at 50 pM. The Taq was Boehringer and Mannheim Taq (1.0 units). This method has been modified from the method outlined by Weisburg *et al.,* (1988).

2.8.2.1 PCR program

The 16S Specific parameters for PCR were: DNA denaturing at 95 °C for 5 min, then 35 cycles of DNA denaturing at 94 °C for 11 sec, 92 °C for 15 sec; then primer annealing at 66 °C for 8 sec, 68 °C for 1 min and extension at 72 °C for 1 min. A final extension of 72 °C for 10 min before it being cooled to 4 °C; samples were then stored at –20° C until required.

2.9 Restriction digestion of DNA

Restriction endonucleases (RE) recognise DNA at specific sites and degrade it by internal cleavage. Enzymes can either cut the DNA to leave no overhanging nucleotides (known as blunt ends), or cut in such a way as to leave an overhang of a few nucleotides at either end of the fragment (sticky ends). As the enzymes are sensitive to contaminants, such as phenol from DNA extraction preparations, the DNA purity will affect the efficiency of the RE. A small sample of the preparation can be digested with an RE known to cut DNA frequently (e.g. *Eco*R1) and is then checked by gel electrophoresis to see if the DNA has been cleaved.

2.9.1 ARDRA analysis

2.9.1.1 Protocol for restriction digestion

The principles for restriction digests are described in Sambrook *et al.*,(1989). The required RE (Boehringer Mannheim, U.K) was added and the mixture incubated overnight for total digestion, at the appropriate temperature, in a temperature controlled water bath, according to the manufacturer's recommendations (Table 2.3).

Table 2.3 List of recommended restriction enzymes (Widmer *et al.*,1998) for use in ARDRA.

Restriction Enzymes	16S PCR products	
	16 Universal	16S Specific
Alu I		
Cfo I	√	√
Dde I		
Hae III	√	√
MspI	√	N/A*
Nde II		
Rsa I	√	√
Taq I		

* Data not available for this restriction enzyme

After digestion, REs were denatured by heating at 65 °C for 15 min and then placed on ice. The restricted DNA fragments were then loaded onto an appropriate agarose gel.

2.9.2 16S Size analysis

Distinct restriction patterns are detected with each RE and strain. Restriction fragments shorter than 90 bp produced by some endonucleases are not resolved well by electrophoresis in 3 % Nusieve. Therefore different restriction profiles are needed to allow identification of bacterial strains. They were determined using band patterns obtained from the following restriction Enzyme Digests: (1) *Rsa I,* (2) *Cfo I,* (3) *Hae III.* This corresponds to results obtained by Widmer *et al.*, (1998), who recommended that a limited range of RE can be used to distinguish between pseudomonad strains using these Pseudomonas-specific 16S primers. The *Hae III* digests were successful in discriminating between isolates, while no obvious differences were apparent with *Rsa I* or *Cfo I.*

2.10. Fingerprint methods

2 10.1. Fingerprint PCR reaction mix

The standard conditions for PCR are noted here; in contrast, specific annealing temperatures or modifications are referred to when mentioned in the text. PCRs were performed in 25 µl total volume. Each solution contained 1 µl template DNA (20–60 ng), 2 µl x 10 PCR reaction buffer (Dynazyme), 1 µl of each primer (50 pM), 0.5 µl of dNTPs mix [10mM conc.] (i.e. 200 µmol of each nucleotide) and 1.0 U [5µ/µl] of Dynazyme. Mineral oil (Sigma) was added to prevent evaporation. DNA amplification was performed in a thermal cycler (TRIO-Thermoblock TBI, Biometra, U.K.). The general parameters for PCR were, denaturing at 94 °C for 7 min, then 30 cycles of DNA denaturing at 94 °C for 1 min. Primer annealing (temperature specific for each reaction) for 1 min, and extension at 72 °C for 8 min. A final extension of 72 °C for 16 min was added to each reaction before it was cooled to 4 °C and maintained. Control reactions were prepared in the same way, but without template DNA. PCR products were observed by electrophoresis on a 2.5 % agarose gel.

2.10.2 ERIC (enterobacterial repetitive intergenic consensus) PCR

ERIC PCR fingerprinting is a method based on short, highly conserved palindromic repeated sequences within the bacterial genome. Methods were developed to classify bacteria on the basis of their genomic DNA patterns and allows for diversity study within a species, at a strain level, by analyzing the specific fingerprint generated from distinct genomes (Rosado *et al.,* 1997). Details of this method are outlined in section 1.2.3.5. The DNA sequence for the ERIC Primers is as follows:

CIRE 1 : 3' **AAG TAA GTG ACT GGG GTG AGC G** 5'
ERIC 2 : 3' **CAC TTA GGG GTC CTC GAA TAT A** 5'

The ERIC oligonucleotides were dissolved in dH_2O, giving a concentration of 50 $\mu m \mu l^{-1}$: the end concentration of each primer in each PCR reaction was 50 pM.

2.10.2.1 ERIC PCR reaction mix

The fingerprint PCR protocol (de Bruijn *et al.,* 1985) was employed with the following modifications: DNA concentration was set at app 50 μg, dNTPs was 200 μM, Primers CIRE 1 and ERIC 2 50 molsl^{-1}. The polymerase was Dynazyme (1 Unit) [5μ/μl]. This method has been modified from that outlined by de Bruijn *et al* (1985).

2.10.2.2 PCR program: ERIC primers for fingerprinting

The general parameters for PCR were: denaturing at 94 °C for 7 min, then 30 cycles of DNA denaturing at 94 °C for 1 min. Primer annealing 52 °C for 1 min, and extension at 72 °C for 8 min. A final extension of 72 °C for 16 min before being cooled to 4 °C and maintained. Total run time was 6 hours 13 min and 12 sec. Control reactions were prepared in the same way but without template DNA. PCR products were observed by electrophoresis on a 1.5 % agarose gel. The fingerprinting gels were run slowly to generate a clear definition of fingerprint patterns.

2.10.3 BOX primers

BOX fingerprint methods work on the same principle as ERIC using different short, highly conserved, palindromic repeated DNA sequences within the bacterial genome.
The DNA sequence for the BOX Primers is as follows:

5' CTA CGG CAA GGC GAT GAC G 3'

2.10.3.1 BOX PCR reaction mix

The ERIC PCR protocol was employed with the following modifications: the DNA concentration was set at approximately 50 μg, dNTPs were at 1.25 mM, BOX primer 100 molsl^{-1}. Taq (Boehringer and Mannheim) at 10 μl^{-1}ml or 2 units of *Taq*. This method has been modified from the method outlined by Versalovic *et al. 1994.*

2.10.3.2 PCR program: BOX primer fingerprinting

The general parameters for PCR were: denaturing at 94 °C for 7 min, then 30-35 cycles of DNA denaturing at 94 °C for 1 min. Primer annealing 55 °C for 1 min, and extension at 72 °C for 8 min. A final extension of 72 °C for 16 min then cooled to 4 °C. Control

reactions were prepared in the same way but without template DNA. PCR products were observed by electrophoresis on a 1.5 % agarose gel. The fingerprinting gels were run slowly to generate a clear definition of fingerprint patterns.

Chapter Three:

Population Size and Composition

3.1 Introduction

This chapter investigates the effects on absolute numbers of culturable bacteria in FYM and sewage treatment plots. The population composition is essentially a simple measure which provides a one-dimensional profile of the soil population against variables within each treatment.

The analysis concentrates on microbial counts of different soil sub-populations using three different media. All of these media measure different soil populations but cumulatively they provide a more holistic profile of different populations within the system. The following media were used:

- 1/10 Tryptone Soya Agar (**TSA**) to select the actively growing heterotrophic bacteria population
- Pseudomonad Selective Agar (**PSA**) to select fluorescent pseudomonads
- MacConkey Agar (**MCA**) to select putative enterobacterial counts

These microbial groups were used to allow an initial assessment of the effect that sampling regimes and/or seasonality have on population structures in the soils sampled.

3.2 Methods

(See Chapter 2 for details of full sampling techniques).

Table 3.1 Sampling times and studies performed

Time	POPULATION SIZE	CATABOLIC DIVERSITY		GENETIC DIVERSITY
		Biolog GN1™	EcoPlates™	
Mar 1997	✓	N/A	N/A	✓
Mar 1998	✓	✓	N/A	✓
Nov 1998	✓	N/A	✓	✓
Mar 1999	N/A	✓	✓	N/A

✓= samples tested at time point; N/A= not available for this time point

3.2.1 Population size

Microbial counts were obtained using standard procedures (Alef & Nanninpieri, 1997). Diluted soil suspensions (from 1 g of sieved bulked soil see 2.2.2) were plated out on a range of selective media. Anti-fungal agents to inhibit growth of fungal populations were not used routinely in this study, as anti-fungal agents may reduce bacterial. With the plating technique, only 0.5-10% of the population can be cultured and anti-fungal agents would only further reduce the recoverable bacteria. Therefore anti-fungal agents were not added to the selection plates. All plates were incubated at 28 °C (section 2.5.2). The following media were used: 1/10 Tryptone Soya Agar (**1/10 TSA**), Pseudomonas Selective Agar (**PSA**), MacConkey Agar (**MCA**).

3.3 Result and discussion

A summary of the results from the microbial counts are presented in Table 3.2, which shows results from both treatments (FYM and Sewage plot). The total viable counts from each plot and each of the selective media at the different sampling times are expressed as cfu gram^{-1} of soil. (section 2.5.2.1).

Table 3.2 Viable counts for sewage and FYM plots at different sampling times on three selective media

	Sewage Plot				FYM Plot		
Sampling Time	**1/10 TSA**	**PSA**	**MCA**		**1/10 TSA**	**PSA**	**MCA**
Mar 1997	$1.4 X10^7$	$2.4 X10^6$	$3.5 X10^5$		$1.6 X10^7$	$4.6x10^6$	$4.2 X10^5$
SE	$7.2 X10^6$	$6.2 X10^5$	$1.3 X10^5$		$6.1 X10^6$	$7.5 X10^5$	$1.5 X10^5$
Mar 1998	$1.4 X10^7$	$2.4 X10^5$	$2.6 X10^5$		$1.6 X10^6$	$4.6 x10^5$	$4.1 X10^6$
SE	$7.2 X10^5$	$6.2 X10^4$	$6.8 X10^4$		$5.8 X10^5$	$7.5 X10^{-6}$	$3.1 X10^6$
Nov1998	$7.0 X10^7$	$5.5 X10^6$	$2.8 X10^4$		$7.5 X10^7$	$3.06x10^6$	$2.1 X10^4$
SE	$7.2 X10^6$	$1.3 X10^5$	$1.3 X10^4$		$1.2 X10^7$	$7.1 X10^5$	$7.5 X10^3$

All counts are expressed as CFU per g of soil Errors are expressed as standard errors of the mean.

3.3.1 1/10 TSA agar

TSA is assumed to select fast-growing heterotrophic bacteria from the soil. The Sewage plots show similar results at each of the sampling times. While both the March 1997 and March 1998 samplings were very similar for FYM , a slightly higher count was reported for the FYM in Nov 1998 (see Fig 3.1). This effect could be due to the sampling regime; the soil samples may have included a 'microbial hotspot', which would account for the slightly higher count. Overall, the result of plate counts of fast-growing heterotrophic bacteria do not show differences in the effect of treatments on the indigenous soil populations. Results indicate that further research is needed to study the dynamics of this broad group of bacteria.

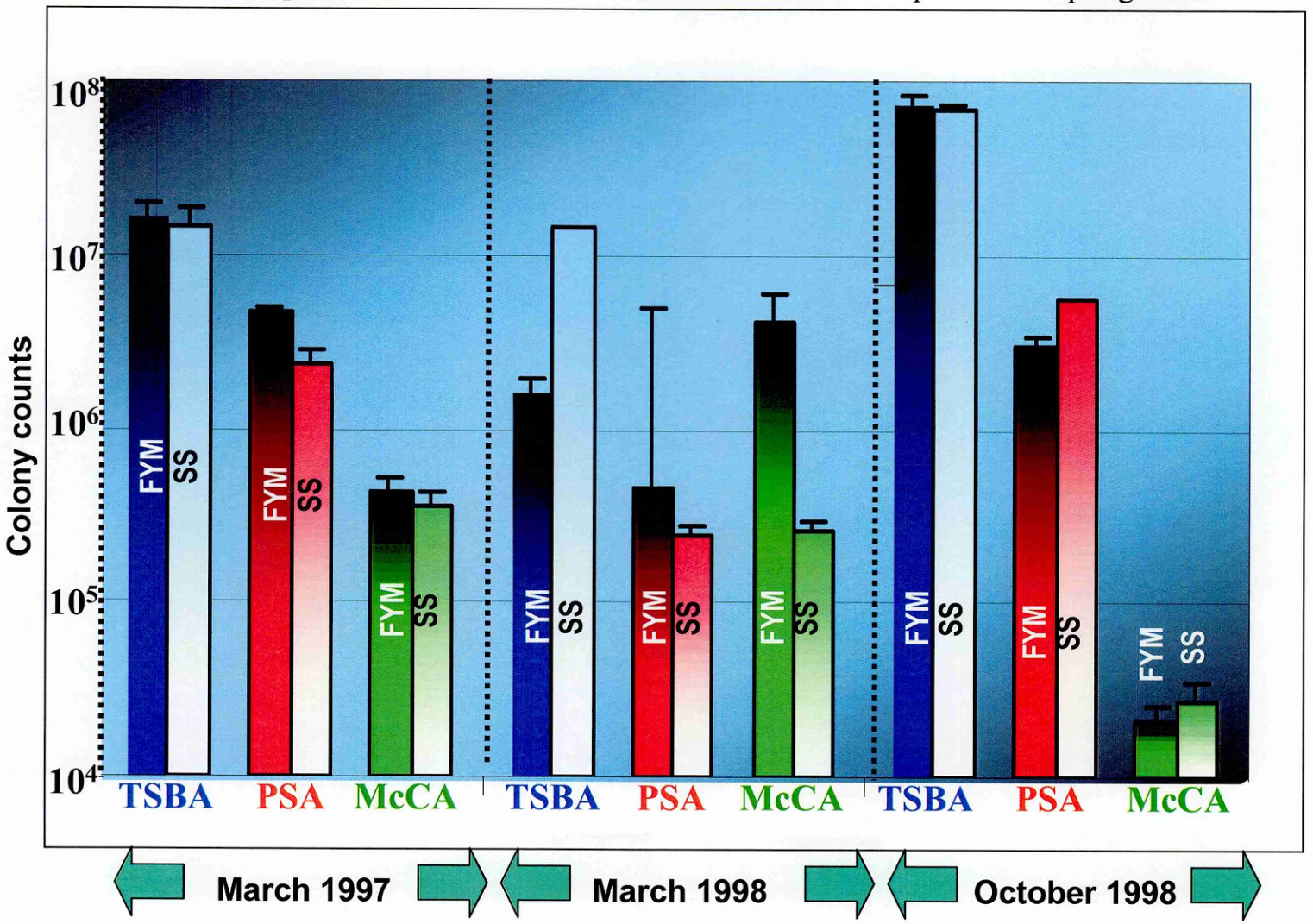
3.3.2 MacConkey agar

The MacConkey agar plates gave similar results for the two treatments, showing no significant effect of the treatment (Figure 3.1). The results for the November sampling were slightly lower than the other two, but not significantly so (Table 3.2). Although there is no difference between the two plots, the results again indicate that the populations obtained in the November sampling were not similar to those obtained in the spring sampling, an indication that there may be some seasonal effect on the number of enterobacteria. The plots have not received any manure application since 1969, so therefore changes in the enterobacteria population are not a result of manure applications, but an effect on the established population of enterobacteria in these plots.

3.3.3 Pseudomonad selective agar

The counts from the PSA media (Figure 3.1) show the difference in pseudomonad counts from both plots. Counts from FYM samples collected during Mar 1997 reveal a significantly higher count of pseudomonads compared to the Sewage plot. This could be an important result because it illustrates differences between the two treated plots. It indicates a possible environmental effect of heavy metal contamination on the pseudomonad group and this has initiated further investigation into their genetic diversity. However, there is only a small difference in March 1998 and this trend was not displayed in the November 1998 sampling, where a significantly higher number of pseudomonads were found in the Sewage plot.

Figure 3.1 Comparison of viable counts on different media at sequential sampling times



This may be attributable to seasonal differences, with higher numbers of pseudomonads found during the Autumn (Table 3.1). It is interesting to note that this potential difference in the pseudomonad population is reflected in the BIOLOG results (Chapter 4). Although it is important to assess population size and overall composition, the value of also studying the genetic diversity of microbial communities is re-enforced.

3.4 Population composition summary

The population size data could be examined in detail by a number of statistical tests. The initial analysis of the data indicates only slight differences between the two plots. However there is an indication of an effect of the time of season (the Autumn sampling differed from the two March samplings). It should be noted that this trend was also reflected in the BIOLOG results (section 4.3). This illustrates the importance in the timing of sampling and highlights the potential problems that might occur with repeated samplings over a period of time. Further study to assess the effect of sampling over a crop growing season would be useful. The results could also suggest that although there was no observable effect on absolute bacterial numbers, there may be a more subtle effect on the genetic composition of the two different soil communities. To prove this, an intensive study of the genetic composition of the pseudomonad population is needed.

3.5 Interpretation of population data

In this section, the data was analysed statistically to reveal any significant trends. The plate count data (Table 3.2) was analysed using the ANOVA test . This allowed analysis of the effect of each medium (Agar plates) on each soil plot (FYM and Sewage) across each sampling time. The Mahalanobis distance was also calculated in order that results could be cross-compared to the Mahalanobis distance results from the BIOLOG data.

3.5.1 Methods

3.5.1.1 Analysis of variance (ANOVA): randomised-block designs

In randomised-block design **ANOVA**, the experimental data are grouped together into sets or blocks. The assumption is that data in the same block will be more similar than data in different blocks. The allocation of the treatments is randomised, independently within each block, in order to maximise the potential identification of the most significant block of data. In the analysis, the aim is to estimate, and remove, the between-block differences, so that the treatment effects can be estimated more precisely.

3.5.1.2 Mahalanobis distance

The Mahalanobis Distance is a frequently used technique to measure the **distance** of a single multivariate observation. It is a calculation based on spatial distance measurement between an imaginary centre of clustered points between treatments.

3.5.2 Result and discussion

This multiple variant analysis allows us to assess the overall influence of different variables in the soil ecosystem. The variables which were assessed were as follows:

(a) The two plots sampled (FYM/Sewage), giving us information on the effect of manure treatment on the actively growing microbial populations

(b) The influence of the sub-populations on the result (i.e. populations isolated on the selective media: 1/10 TSA; PSA; MCA)

(c) The sampling times (Mar 97, Mar 98, Nov 98) which give some information on the seasonal effect on the populations

The results are summarised in Table 3.2, showing a ranking of the effect of the media of two key variables, namely:

1. Difference between plots
2. Differences in sampling times

Table 3.3 ANOVA analysis showing the ranking of each sampling, with the order of means on each selection media

(a) ANOVA grand mean results. This shows the effect that each mean of each sampling has on each medium.

Media	PSA		MCA		TSA
S98 Nov	5.33 x10^6	F97	4.33 x10^5	F98 Nov	7.00 x10^7
F97	4.77 x10^6	F98 Mar	3.77 x10^5	F97	2.00 x10^7
S98 Mar	2.87 x10^6	S97	3.67 x10^5	S97	1.00 x10^7
S97	2.67 x10^6	S98 Mar	2.63 x10^5	S98 Nov	1.00 x10^7
F98 Mar	4.77 x10^5	S98 Nov	2.83 x10^4	F98 Mar	2.00 x10^6
F98 Nov	2.87 x10^5	F98 Nov	2.33 x10^4	S98 Mar	1.00 x10^6
Grand means	**2.73** x10^6		**2.49** x10^5		**1.84** x10^7

(b) Analysis of variance. This shows the overall effect of the bacterial counts have on the results. This is obtained by running an analysis on all media and all sample points together.

Analysis of variance			
	PSA	**MCA**	**TSA**
F5,12	19.08	7.92	92.52
p=	<0.001	0.02	<0.001
l.s.d (5%)	1479616	198299.9	8063679

This analysis shows that there was no one factor with a strong influence on the results. Therefore an effect on the microbial populations cannot be connected to one of the variables set out, but there may be *interactions* between more than one factor which influence the results. The effect on the population size (if present) is not a simple one-dimensional effect, but is a combination of more than one of the outlined variables. This indicates that the influencing factors may be hidden in the large amount of data that has been analysed. This suggests that there is a complex interaction between variables, which must be examined in different combinations to leverage further insight into the system.

3.5.3 Conclusion

Overall, results from these analyses show that there is as much difference between sampling points as there is between plots, so results are inconclusive. The relative importance of populations selected on each medium is unclear. Therefore, it is impossible to draw any substantive conclusions looking at each discrete population set alone, particularly when differences could be accounted for by random fluctuations.

3.6 Population size variables

As outlined above, there are a number of contributing variables in this investigation which have led to the complex results. Since no one variable listed in 3.5.2 was found to have a direct, overall effect on these soil systems, a contribution from more than one variable has to be considered. The results show no definite influences were obtained, except for the combination of the two variables: the pseudomonads group and putative enterobacteria. When the PSA counts and MCA plate count data were analysed together (section 3.5.2), some intriguing results were produced (Table 3.4).

3.6.1 Pseudomonad and enterobacterial influence

The aim of the analysis was to investigate the influence of the two different treatments on the sizes of the microbial populations and to assess the effect of both the sampling regime and timing on the overall result. This is because the analysis of each medium separately did not provide a discernible trend between different soil microbial populations. The data sets from both media (PSA +MCA) were analysed together, using the Mahalanobis distances, and the following results were calculated. (The two media were analysed as one influencing factor and Table 3.4 shows the sampling effect between different soil populations at different sampling times). The Mahalanobis distances are used to explain the relationship between different time points and the two plots.

Table 3.4 Mahalanobis distances from joint analysis of PSA and MCA data

*** Inter-group distances *** Mahalanobis distances						
FYM Mar '97	0					
Sewage Mar '97	2.289	0				
FYM Mar '98	3.119	5.407	0			
Sewage Mar '98	2.951	5.184	1.017	0		
FYM Nov '98	4.737	3.903	6.942	7.29	0	
Sewage Nov '97	3.103	4.111	3.984	4.593	3.262	0
	FYM Mar '97	Sewage Mar '97	FYM Mar '98	Sewage Mar '98	FYM Nov '98	Sewage Nov '97

3.6.2 Variables summary

The analysis shows that there are subtle differences between the two plots. The similarity between the soil plots taken in the two March samplings is pronounced, whereas the November sampling indicates differences in soil sub-populations, which implies a seasonal effect. There is some consistency in the effect of season on FYM, whereas the effect on Sewage is random. This could suggest that variations between years, but within seasons, are narrower because of a number of environmental factors i.e. soil moisture, carbon content and temperature. Therefore, the significance of the sampling regime on soil population appears to be more significant than differences between soil treatments. Clearly such factors vary with the season. The data collected to date suggests that a more formal investigation of seasonal variables on soil environmental populations is necessary to shape more definitive conclusions.

3.7 Conclusions

There are limitations to the measurements of microbial populations which stand in the way of any substantive conclusions. Essentially, the quantitative data (Total Viable Count or 'headcount') of the selected culturable sub-populations lacks the breadth and detail necessary to determine the full effects of metal contamination on soil population structure. Indeed, while it shows both plots have very active and diverse heterotrophic populations, it fails to reveal any depth of information on structure, particularly what is

happening to the pseudomonad population. In contrast, the study of the *Rhizobium* population in metal contaminated sites (Hirsch, 1986) showed the effects of metal contamination on the Rhizobium community in metal contaminated plots, which was drastically reduced both in size and genetic diversity.

There are three possible reasons which may account for population levels of pseudomonads remaining constant. One is that there is greater **bacterial redundancy** in the pseudomonad population (Yin *et al.*, 2000), which means that the amount and variety of functions carried out by pseudomonads are much less than the available 'capacity' of the population to carry out those tasks. The second is that pseudomonads may be genetically capable of 'switching' over to another function in order to carry out existing activities within the soil. The third is the inherent metal resistance which confers the pseudomonads population with the natural ability to survive in soils with elevated levels of heavy metals (Bruce 1997). In all cases, it is necessary to profile **genetic diversity** in order to appreciate the effect that long term heavy metal contamination has on the pseudomonad population.

The **sampling regime** is a significant determinant of the quality of the output. Ecologists tend to advocate numerous samples at different times (Begon *et al.,* 1986, Hawksworth 1995), rather than single large samples, which is common practice in microbiology. While a very large and diverse population will be captured in a single sample (highlighting the most abundant species at that time), each subsequent sampling increases the chances of surveying the **richness of the ecosystem** as well. (The benefits of this approach are discussed in more detail in section 4 3.1.1).

There would appear to be little value therefore in increasing the volume and frequency of samples taken, unless genetic profiling is introduced as a new measurement technique. This is simply because population data derived from viable counts to date could easily conceal more significant changes in the pseudomonad population itself, among them genetic re-configuration and changes in distribution of functional tasks. This implies a need for a more detailed **genetic profiling** of soil sub-populations in order to understand the broader consequences of environmental contamination on soil microbial functions at a genetic level.

Chapter Four:
Catabolic Diversity

4.1 Introduction

This chapter describes the techniques which have been developed to study microbial catabolic diversity. The work involves analysis of soil samples taken at multiple sampling times to show temporal changes in diversity. The multi-sampling approach is advocated by ecologists studying plant and animal populations (Begon *et al*, 1986, Hawksworth 1995), but is in contrast to microbiologists who typically employ a large one-off sampling approach. The sampling strategy used here allows investigation of population structure as influenced by two key variables: season and plot treatment.

4.1.2 Catabolic profiles

The study of catabolic functions of soil has been used by numerous groups to provide an indication of the functional ability of the soil (Garland *et al.*,1991; Zak *et al.*,1986, O'Flaherty *et al.*, 2001). The BIOLOG GN1™ technique allows a crude estimation of the functionality of the soil: it is defined by its ability to utilise a range of carbon substrates giving a catabolic profile of the soil. Despite the limitations of this methodology, it attains a measure of catabolic diversity, allowing us to compare the culturable, fast growing, heterotrophic population of soil biota in different samples.

The fundamental problem facing analysis of soil microbial populations is the sheer number of unknown species. But because this study concentrates on microbial populations that can be cultured, the use of BIOLOG™, even in this case, is justified. The BIOLOG GN1™ system was designed to characterise microbial catabolic diversity by determining the range of carbon substrates which micro-organisms are capable of metabolising. Previous studies have shown that substrate utilisation patterns for whole communities generated by BIOLOG™, differ between diverse habitats (Garland and Mills 1986, Garland 1987, Grayston *et al.*, 1998). However, any measure that is applied to soil communities has to acknowledge the probability that data captured is only representative of a fraction of the total soil population that can grow in the BIOLOG GN1™ plates.

The Average Colour Well Development (ACWD) is a simple measurement that is the mean of the absorbance reading of all the wells on the BIOLOG GN1™ at a single time point (Schneider and de Bruijn, 1996). This measurement was commonly used when the BIOLOG GN1™ system was first applied to soil but in more recent times more complicated analyses have been applied, and analysis programmes developed to include them.

4.1.3 Alternative diversity measurements

There are a range of ecological indices which are mathematically calculated to give an expression of measures to soil catabolic diversity. These can be used to describe how diverse a given soil ecosystem is. Given that the variation in these mathematical analyses produces different results depending on the question been posed, it is important to articulate a clear rationale for the selection of one ecological index over another. There is as yet no discernible consensus around the use of these indices; there are advantages and disadvantages depending on the particular experimental application and the questions being asked. The reason the scientific community sustains so many competing indices is that ecologists studying different ecosystems analyse and define diversity in their own terms to facilitate their particular area of research. This profusion of indices is discussed in a book dedicated to this question (Magurran 1988).

Table 4.1 Discrimination ability of a range of diversity measurements

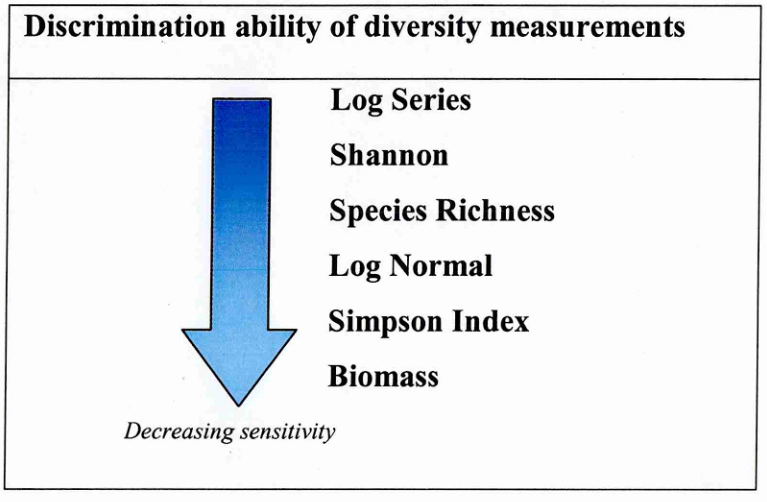


Table 4.1 displays a small range of diversity indices which are commonly used, showing the potential discrimination ability of the indices in relation to each other (Magurran, 1988). The debate on the difference in diversity measurements is, however, beyond the scope of this thesis. From the enormous range of measurements available, this study has selected two principal indices which are frequently used and referenced by microbiologists. They are the Shannon Index (Shannon, 1949) and the Simpson Index (Simpson *et al.*,1949). Both are based on the proportional abundance of species and seek to crystallise **'richness'** and **'evenness'** into a single figure. 'Richness' is defined as the number of species within a given system, while 'evenness' is defined as the way the number of these species are distributed within a given system. The calculation of a single value is useful, as it can be used to compare diversity within different systems.

The Shannon Index: This assumes that individuals are randomly sampled from an 'indefinitely large' population and that all species are represented in the sample. The value of the Shannon diversity index is usually found between 1.5 and 3.5 and if calculated for a number of samples the index will be distributed normally. It is affected by rare species (i.e. species richness) so, logically, gives an indication of species richness. The Shannon index is not biased when the sample size is small and should not increase as sample size increases.

Simpson Index: This is a dominance measure, since it is weighted towards the abundance of the commonest species, rather than proving a measure of species richness. This index is sensitive to changes in the abundance of the commonest species (i.e. evenness/dominance) and is therefore influenced by sample size, tending to increase as sample size increases.

Both indices are in common use and provide two different approaches to understanding diversity. The **Simpson Index** is heavily weighted towards the most abundant species in the sample while being less sensitive to species richness. It has been shown to be less sensitive to inter-site differences than the Shannon Index (Magurran, 1988). The Simpson index has proved useful at detecting shifts in dominance. The Shannon Index

is widely used in pollution monitoring, to indicate the species richness of a given system, while the Simpson Index acts as a measure of either species diversity, or the diversity of resources. Although the two indices operate differently, when applied together, they can provide information on both species dominance and richness in each community structure, giving a simple value which allows direct comparison (Kempton, 1979).

The indices were conceived originally to analyse known plant or animal species in an environmental sample. However the same logic can be applied using the same methodology to analyse the utilisation rate of a range of carbon substrates. In this way, each separate well (containing a different carbon substrate) in the BIOLOG GN1™ plates is considered to be a single species, although there are clearly many bacterial species in each well. Both indices can provide insights into the diversity in microbial communities and their catabolic abilities. The BIOLOG GN1™ system provides an understanding of the **dominance** of particular function in the community, as well as a profile of the different **richness** within each system. The application of the catabolic diversity profiles can be justified by using these indices to display **niche width** as illustrated by Magurran (1988). The latter is described as an abundance measurement in a system, comparing the number of individuals either eating each type of food, living in each sort of habitat, or adopting each kind of behaviour. Therefore, by observing the utilisation of carbon sources, these can be used as an indication of niche width, and from this, an indication of diversity can be derived. Although their utility as definitive measures is limited, these indices provide an indication of microbial diversity.

Reviewing these concepts was vital for the design and development of a set of Genstat Programs (Appendix A: "BIOLOG GN1™ 96" and "EcoPlates™ 32"), which calculate the indices discussed and provide a useful aggregation of otherwise divergent measures. However, having acknowledged their limitations, the diversity indices can provide a reliable profile of activity within different soil environments. The application of BIOLOG GN1™ systems (i.e. both the BIOLOG GN1™ and EcoPlates™ well systems) as supplied by BIOLOG GN1™ Inc were assessed. The reliability of the measures by the BIOLOG GN1™ systems could be improved by including more

environmentally significant substrate ranges as advocated by Campbell *et al.*, (1997). However, this study used the commercially available product BIOLOG GN1™ and EcoPlates™, although an expanded carbon substrate range would be considered for further studies.

4.2 Materials and methods

The experimental set-up of the BIOLOG GN1™ (also referred to as GN Microplate™ in the literature) and EcoPlates™ systems have been standardised and are described in detail in 2.4.2.

Table 4.2 Different sampling times for each of the diversity indices in a research programme carried out over a 18 month period

Sampling Times	CATABOLIC DIVERSITY	
	BIOLOG™	EcoPlates™
Mar 1998	+	N/A
Oct 1998	N/A	+
Mar 1999	+	+

N/A = data not available

+ = data collected for this time point

The objective was to observe spatial and temporal variations within soil microbial populations, concentrating on the three diversity measurements. It is important to recognise the inherent variability in any given soil community, implying a low probability of deriving the same result at each sampling point. However, the central focus of the research is not to obtain identical results, but to observe if trends within data may be relevant.

4.2.1 Methods of analysis

Computerised statistical programs were developed to analyse data for both BIOLOG GN1™ and EcoPlates™ (Appendix A). The program was designed to calculate both ecological diversity (i.e. that which is observable within each substrate) and statistical measurements. Data was obtained at different time points (0, 12 , 24, 36, **48**, 52, 64, 72, 96 hour). By using utilisation curves (i.e. the rate at which the carbon sources were oxidised) the sample time of 48 hours was chosen as the standard time for the analysis. The interpretation of BIOLOG GN1™ data can be notoriously complex, so a simple approach has been taken and results are reported in two ways:

1) Ecological diversity indices (4.1.3)
2) Other statistical analyses (4.2.1).

The use of diversity indices has, due to their very nature, relies on assumptions about data (i.e. the calculations within indices are made with preconceived ideas of the population being tested). The approach in this thesis is different given that data is reviewed in a purely mathematical way and no preconceived assumptions are made about it. Instead, raw data is taken and analysed. There are a number of different mathematical approaches to statistical analysis and a limited number of relevant approaches are presented. The main components of each approach is outlined below.

4.2.1.1 Principal components analysis

Principal **C**omponents **A**nalysis (**PCA**) is method which calculates and then ranks the data, in order (or factors) affecting it. These are expressed as components (or variants). Each principal component (PC) accounts for a variation in data. By establishing orthogonal linear combinations amongst a set of varieties, PCA maximises visible variations, displaying most of the original variability across a smaller number of dimensions by using the sum of squares (Genstat Manual,1988 and Appendix A).

PCS components are then labelled in a range from component 1 (PC1) to component n (PCn). The first component represents the highest percentage of data variation, i.e. identifies the most significant factor influencing the data; the next component represents

the next highest percentage effect, and so forth. The percentage of variation shows how important the different components are in relation to each other. An indicative analysis typically identifies the variations shown in the table below:

Table 4.3 Typical percentage variation with experimental BIOLOG GN1™ data

Principal Component	% Variation
PC1	40-60
PC2	15-30
PC3	5-10
PC4	2-10
PCn	n-n

The largest variation in results is described by component 1 (PC1), which identifies factors having the largest significance. PC2 includes factors having less influence, although these can contribute up to 30% of the variation, with PC3 describing a lesser influence, and so on. The *causes* of these influences on the components are not identified by this analysis.

The PCs are calculated by the **P**rincipal **C**omponents **A**nalysis program in Genstat and data is presented in the form of an ordination plot, which represents the factor and its position of influence on the data. These co-ordinates can then be plotted against each other. In general, the results are represented in a two dimensional diagram which, to all intents and purposes, resembles a conventional XY plot. It is important to stress that each PC represents one of the axes (and as stated earlier) has a percentage of influence on the data. Therefore each axis is weighted differently (with one having greater influence in determining the data position than the other). Despite this departure from conventional graphing techniques, displaying the results in PCA graph format allows ready interpretation of key trends within the data.

4.2.1.2 Procrustes rotation

To try to understand the variation between all the sampling times, it was necessary to draw up comparisons between all the sampling points. An analysis called **Procrustes Rotation** was used; this is a mathematical calculation allowing comparison of two data sets obtained by PCA analysis (i.e. the components of the PCA). The basic principle is to take the PCA scores of two analysis and physically impose both data sets on top of each other to find the 'best match' on the graph, via the rotated positioning of the points. (Genstat Newsletter 1988).

One set of data is held in a fixed position and the 2nd set is rotated in the different spheres (i.e. planes) to attain the best match. This methodology is known as the 'classical pairwise Procrustes matching' (Genstat Manual, 1988). It is a robust method which allows comparisons between sampling points and an investigation of the variability inherent in BIOLOG GN1™ data analyses. It also permits analysis of the data from an entirely different perspective, namely a comparison of the catabolic profiles obtained with the different BIOLOG GN1™ systems and sampling times. It facilitates a reliable and objective evaluation of both systems, by isolating the effects which data from different sample time points otherwise have on the interpretation of catabolic diversity.

4.2.1.3 Canonical correlation analysis

Canonical Correlation Analysis (CVA), (Digby & Kempton, 1987) is a method which correlates PCA scores into a statistical mean. In simple terms, this analysis calculates the mean of the replica PC scores for each treatment and provides one set of co-ordinates. The CVA scores are represented on a diagram similar to the PCA graphs (see 4.2.1.1). However, the PCA graphs were routinely used to allow the variation between the sampling replicas to be seen. The CVA were performed to calculate the inter-group-mean i.e. Mahalanobis distances which is an another statistical technique used.

4.2.1.4 Mahalanobis distance

The Mahalanobis distance is a technique used to measure the **distance** of a single multivariate observation from an imaginary centre of clustered points (Manly, 1994).

The Mahalanobis distance is a measurement technique which can be applied to any multivariate observations and allows for important cross-referencing of results derived from different methods. One constraint has been identified with this measurement: the data obtained from the genetic diversity study (section 9.5) data was complex, and the Mahalanobis distance could not be calculated. However, the similarity matrix calculation (which is also a distance measurement) is directly comparable (D. *Murray* personal communication, 1999, 2000) and thus it was used in the genetic study (section 9.5.4).

4.3 Results

4.3.1 Ecological diversity

The methodology used was intended to focus on two key areas: species richness and species dominance in order to show changes in catabolic diversity and to observe if heavy metal contamination affected any aspect of soil microbial communities within each plot.

Table 4.4 Different methods used to calculate diversity on BIOLOG ™ data

INDEX	SIMPSON		SHANNON		AWCD	
FYM mar 98[B]	36.85	*1.12*	3.90	*0.02*	0.37	*0.01*
Sewage mar 98[B]	40.48	*1.74*	3.98	*0.04*	0.43	*0.01*
FYM Nov 98[E]	15.69	*0.95*	2.92	*0.04*	0.66	*0.02*
Sewage Nov 98[E]	15.62	*1.08*	2.91	*0.07*	0.58	*0.05*
FYM Mar 99[B]	36.00	*3.87*	3.85	*0.08*	0.26	*0.00*
Sewage Mar 99[B]	30.11	*2.39*	3.81	*0.07*	0.25	*0.00*
FYM Mar 99[E1]	9.00	*0.79*	2.41	*0.10*	0.12	*0.03*
Sewage Mar 99[E1]	11.57	*0.21*	2.64	*0.02*	0.28	*0.01*
FYM Mar 99[E2]	11.48	*0.94*	2.61	*0.08*	0.23	*0.03*
Sewage Mar 99[E2]	11.87	*0.91*	2.65	*0.07*	0.25	*0.03*

***B** indicates the 96 well BIOLOG ™ plates* **+/- Standard Errors in *italic text* are to 3 significant figures.** ***E** indicates the 32 well EcoPlates™*

4.3.1.1 Substrate richness

Substrate richness appeared to be similar in the two plots, as indicated by the Shannon Index (Table 4.4). However, this apparent consistency may be due to either one, or several, of three factors:

1) The method is insufficiently sensitive to pick up evidence of rarer species
2) The limited range of substrates (or the actual composition of the carbon sources) impeded the ability to discern species richness
3) There is no overall difference in the two plots. Potential solutions to test these soils further could be to increase the total number of substrates and/or add more environmentally significant carbon substrates (Campbell *et al.,* 1998). There are limitations in looking at species diversity, using either the Shannon or Gini indices. The Gini index is simply an inverse calculation of Shannon, which is also widely cited in the literature (hence mentioned here). At any rate, the correlation between the two indices is, at 0.99, exceptionally high, so the discussion will concentrate only one index , i.e. the Shannon Index.

The Shannon index showed no differences between any two plots. The BIOLOG GN1™ system for both March 1998 and March 1999 registers species diversity at between **3.8 and 3.9**, while the EcoPlates™ gave a lower index value, between **2.4-2.9** (Table 4.4). This highlighted an interesting finding, namely that any species richness calculation was directly affected by the number of substrates used. To infer anything reliably about species richness in a microbial community, it would appear essential to investigate the effect of increasing the number of carbon substrates used.

4.3.1.2 Carbon-Sources dominance

Carbon-sources dominance, assessed by the Simpson index, showed differences between soil plots (Table 4.4) This change itself was variable because Species Dominance was dependent on sampling times, hence making any assessment of species dominance difficult to interpret . The index was clearly influenced by the sample size, making it impossible to compare BIOLOG GN1™ and EcoPlates™ results.

The BIOLOG GN1™ methodology indicated differences in species dominance between both plots, although inconsistently. This observation could be attributable to the fact that what was been observed was the constantly shifting dominance of individual bacteria clusters in the soil, over time and between samples. However, the potential population shifts can be seen as an important consideration, because unlike in the animal or plant kingdoms, where species dominance is largely established, it would appear that in communities of soil bacteria relationships are more fluid. The dominant bacteria could be transient and simply reflect differences in soil environments at the time of sampling, rather than any more significant indication of status as consistently dominant species. This result is important because it questions the application of general ecological principles to the study of soil microbial populations. This question has been raised at a number of relevant conferences, but the problem has yet to be addressed.

4.3.2 Principal components analysis

The selected data point of 48 hours was selected from the utilisation curves (data not shown) and 48 hours was used in all analysis to aid comparisons between sampling (see Table 4.2. for the multiple sampling strategy).The analysis used the BIOLOG GN1™ and EcoPlates™ Genstat Program (Appendix A). The percentage PCA scores which were obtained are shown in Table 4.6.

Table 4.5 Percentage variation of PCA in the first 5 dimensions

System/Time	PC1	PC2	PC3	PC4	PC5
Mar 98 BIOLOG™	35.63	27.3	15.94	12.3	8.84
Nov 98 EcoPlates™	32.57	25.12	21.86	14.57	5.88
Mar 99 BIOLOG™	42.46	18.04	17.49	13.06	8.95
Mar 99 EcoPlates™ (A)	40.84	28.46	16.23	10.58	3.89
Mar 99 EcoPlates™ (B)	44.65	28.7	14.94	6.35	5.37

The PCA results showed there were consistent differences between the FYM and Sewage plots. A representation of these plots can be seen in Figure 4.1; all other PCA results can be seen in Appendix B. These results show the general findings of the PCA analysis. Figure 4.1 highlights the differences between the two plots: with the 96 substrate BIOLOG GN1™ plates, the three replicates for the sewage plot are strongly correlated and fall into a tight cluster together. The FYM replicates fall within the same section of component one, but form a looser cluster than the sewage plot. These results are similar at each of the sampling times (i.e. Mar 97, Nov 98, Mar 98, Mar 99, see Appendix B).

The PCA analysis shows that the two plots are different. The key finding is that the samples are clustered together into different halves of the plot. It shows that FYM has more variability between replicates than the Sewage plots. These trends are always consistent at each of the sampling times (See Appendix B).

The comparison between the EcoPlates™ and BIOLOG GN1™ show that the EcoPlates™ loosely show the same general trend as BIOLOG™, but there is much more variability with the replicates on the EcoPlates™ (Figure 4.2). However, it is important to highlight that the mean of the replicates points to the same trends, even if there is more inherent variability between replicates in the EcoPlates™ than between plots (MAFF SP 0120, 1999). This, again, shows how the sensitivity of the BIOLOG GN1™ method is affected by the number and range of substrates used in this test.

Fig 4.1 A PCA plot for FYM and Sewage (Mar '98), using BIOLOG

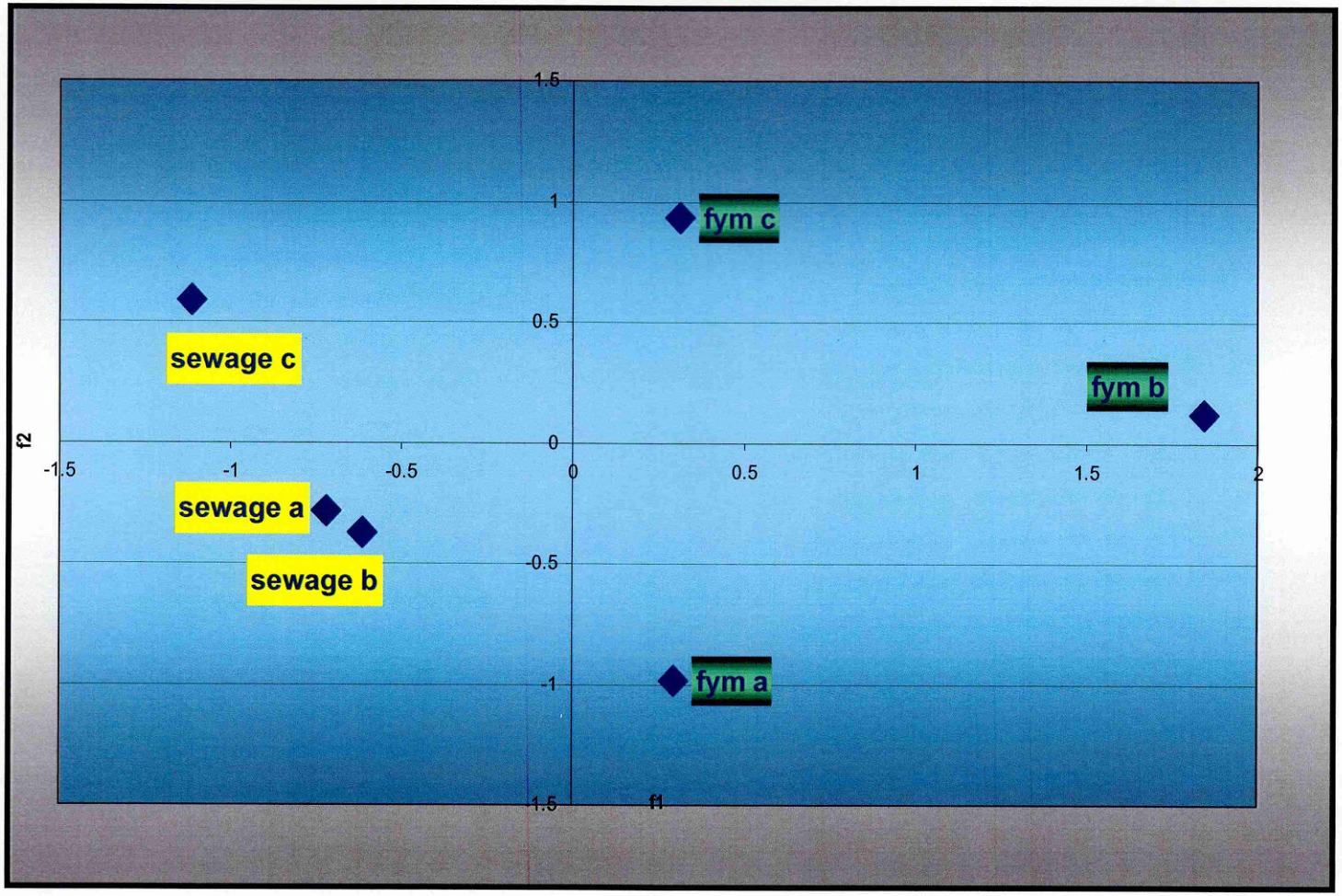
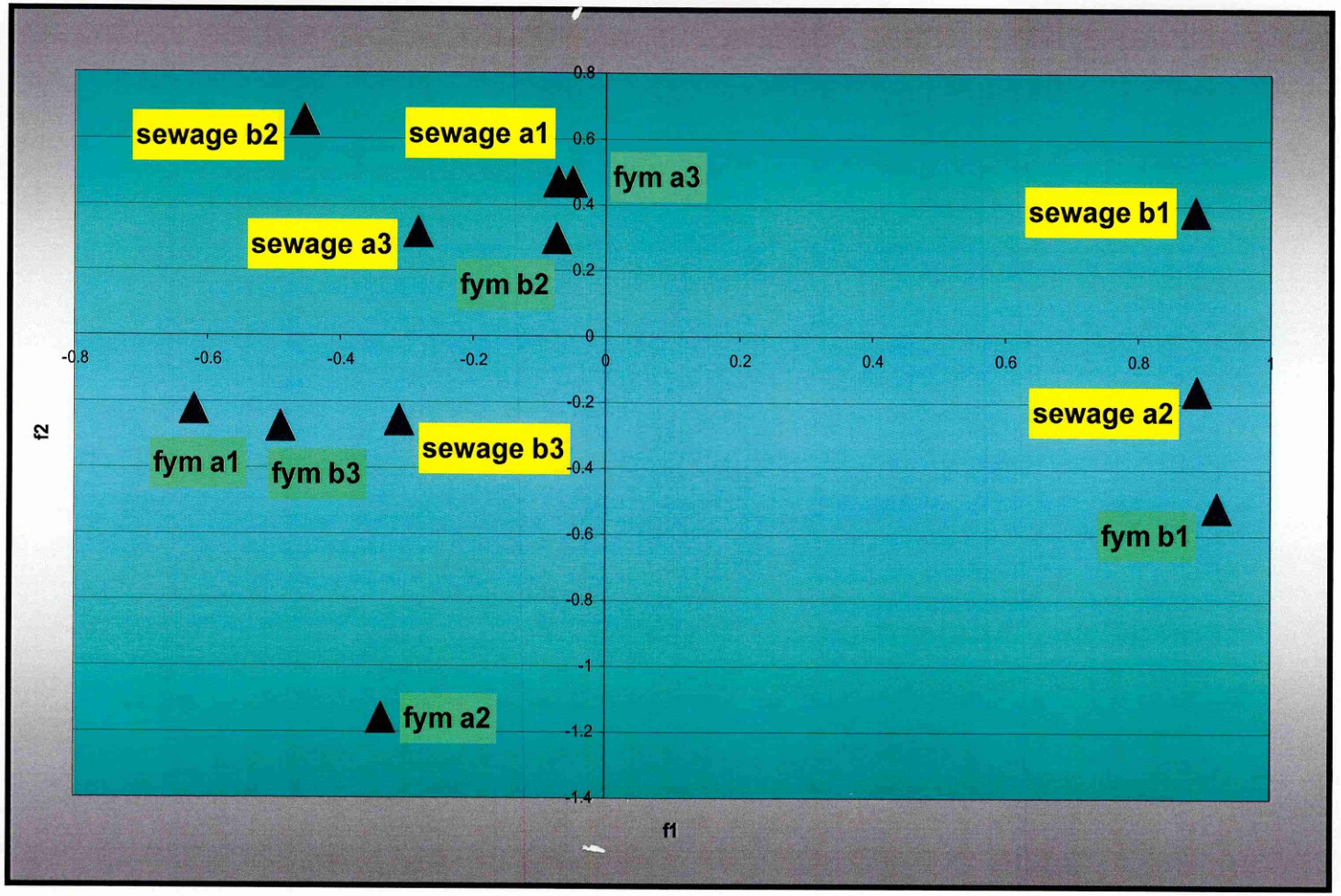


Fig 4.2 Variability in EcoMicroplates (Mar '99)



4.3.3 Procrustes analysis rotation

The result shown in Table 4.7 allowed us to investigate the reproducibility of the BIOLOG GN1™ system (Appendix B). It was also used to compare the two different BIOLOG GN1™ systems, BIOLOG GN1™ and EcoPlates™

Table 4.6 Degree of matching obtained from Procrustes Rotation

Samples	Rating
Seasonal Variation	
BIOLOG ™ Mar 98 to Mar '99	**Reasonable**
EcoPlates™ Nov 98 to Mar 99	**No match**
EcoPlates™ Mar [1] 99 to Mar [2] '99	**No match**
Validation of the Systems	
BIOLOG GN1™ Mar 98 to EcoPlates™ Nov 98	**No match**
BIOLOG GN1™ Mar 98 to EcoPlates™ Mar [2] 99	**Reasonable**
BIOLOG GN1™ Mar 99 to EcoPlates™ Nov 98	**Good**
BIOLOG GN1™ Mar 99 to EcoPlates™ Mar [2] 99	**Good**

4.3.3.1 Validation of EcoPlates™

(1) Comparing EcoPlates™ Mar 1998 to Mar 99

The results from this comparison show that the was no similarity between the results obtained from these two sampling points, indicating that the seasonal effect on the EcoPlates™ cannot be assessed.

(2) Comparing EcoPlates™ replicates of Mar 1999

The comparison between two EcoPlates™ from the sampling time point and plots were investigated. The Procrustes rotation was used to examine the PCA scores: it showed that there was no match from either the FYM or Sewage plots, both in terms of the plots and the replicate plates from within each plot. The results suggest that there was no consistency with the reps from both the FYM plot and the Sewage plot. The replicate plates give dissimilar results.

4.3.3.2 Validation of BIOLOG GN1™

Comparing BIOLOG GN1™ at two sampling times, Mar 1998 and Mar 1999

The Procrustes Rotation results from both March 98 and March 99, showed that the Sewage plots cluster together, while (once again) the FYM Plots cluster together, but there is more variation within the reps. The result indicated that the BIOLOG GN1™ system was a robust system and was able to identify gross differences between the heterotrophic populations and also, more importantly, the result was reproducible. The results presented here validated the uses of the BIOLOG GN1™ system in assessing the gross catabolic diversity in the soil population.

4.3.3.3 BIOLOG GN1™ & EcoPlates™ comparison

(1) Comparing EcoPlates™ 1998 Nov to BIOLOG GN1™ Mar 1998

The Procrustes rotation found only a reasonable match between these two samplings. However the same trends were still seen i.e. Sewage plots tend to cluster together in one quadrant and the FYM plots in another.

(2) Comparing EcoPlates™ 1998 to BIOLOG, Mar 1999

Procrustes showed a good fit with EcoPlates™ '99 (rep 2) and BIOLOG GN1™ '99, perhaps surprisingly, as no fit was found between the EcoPlates™. The Sewage plot is a very good fit, i.e. the positions on the graphs were almost identical. The FYM plots indicate more varied results, although this may simply reflect the previous observation that there was, in any case, more variability within reps from FYM compared to Sewage Plots.

4.3.4 Mahalanobis distance

The differences between the plots were calculated and expressed via the Mahalanobis distance, the results of which are shown below.

Table 4.7 Mahalanobis distance showing inter-group distances between the two Woburn Plots from BIOLOG GN1™ data

Time points	Mahalanobis distance*
BIOLOG GN1™ Mar 98	0.90
BIOLOG GN1™ Mar 99	0.60
EcoPlates™ Oct 98	0.68
EcoPlates™ Mar99 (b)	0.54
EcoPlates™ Mar99 (a)	1.80

* Significant level set at 95%

The Mahalanobis distance conceptually calculates the distance between clusters, by using mean differences between treatments. CVA were performed to calculate the inter-group-mean Mahalanobis distances between the sites. The differences between the two plots showed in Table 4.6. There was variation between the length of distances between the two plots i.e. between 0.54 and 1.80, however there was a consistent difference, indicating that the two plots were different.

The distance results confirmed the conclusions drawn from the PCA profiles, i.e. that the plots were different and that the trend was similar at each sampling point. Using the BIOLOG GN1™ systems, changes can be detected in the catabolic profiles within the two plots.

4.4 Discussion

The ecological indices were designed to reflect the population structure in the plant and animal kingdoms, and have, therefore a number of underlying assumptions relating to them. The question that this research poses is: can these populations assumptions be applied to constantly changing microbial communities? The concept of species diversity and population structure, as it applies to the animal and plant kingdoms, cannot validly be transferred to microbial communities. Although these measurements provide a useful

snapshot of the environment at a precise sampling time, their relevance and application is uncertain. It is probably more appropriate to rely on a more mathematical approach, which is not dependent on such assumptions to support its conclusions.

There is limited knowledge of species diversities in the soil environment and unlike the other kingdoms, microbial communities (or at least the ones that we have knowledge of) are very diverse, and classification and re-classification of bacterial species is still on-going. Further studies need to be completed before any realistic attempt can be made to either apply the ecological measurements mentioned above. Alternatively, perhaps as a result of newer research, new formulas and indices will have to be developed to reflect diverse and changing microbial communities. Little is known about the depth of diversity among microbial species in soil and new approaches towards mathematical analyses may be needed to interpret the data. After all, one of the features which separates the microbial kingdom from any other is the importance of functional redundancy *(Yin* et al., *2001)*, resulting in species redundancy within their ecosystems.

Bacterial functional redundancy is defined as the underlying population of bacteria in the community, which can be selected to replace a lost important bacterial group, or function, in the system (Yin *et al.,* 2001). It has been shown that there is a more obvious relationship between bacterial communities and the re-growth of plant structures in stressed soils, showing that bacteria are an important factor in the restoration of biological functionality (*Yin et al.,* 2001). These changes in bacterial populations, or functions, may be observed using the BIOLOG GN1™ system. Hence BIOLOG GN1™ has a purpose in monitoring changes in population function with changes in crop, although we question its sensitivity in showing shifts in population structure, due to subtle levels of heavy metal contamination.

4.5 Conclusions

Changes in microbial diversity will not always correspond to changes in functional redundancy, and vice versa . Hence, when evaluating biological diversity, it is important to distinguish between these two parameters. While species diversity indices can provide a broad measure of biological diversity, whether they reliably indicate

catabolic diversity is a matter which may only be settled by further in depth study. However, the BIOLOG GN1™ systems *did* detect changes in catabolic profiles between the FYM plot and Sewage plot.

The BIOLOG GN1™ was shown to be a robust system in assessing gross changes in the catabolic profiles in the soils tested. However the results highlight problems with the EcoPlates™ system (due to the limitations of the substrates range), and indicate that it is not suitable in detecting the subtle differences present in this experiment. Further work is needed to investigate the EcoPlates™ system before it can be validated. It also shows that although there is a difference, which varies according by sampling time, this could be attributable to either the possible seasonal effect on this system, or otherwise the inherent variability of the BIOLOG GN1™ system itself.
Despite these qualifications, an important and probably highly significant finding is that there is a discernible difference between two heterotrophic communities in the two soil plots.

The results from the catabolic profiles study indicate that there are differences between the actively growing heterotrophic populations in these two plots. The differences were highlighted suggesting that the BIOLOG GN1™ system had the ability to discriminate and display these differences. Therefore, it could indicate its value as a possible indicator of environmental bacteria shifts. However, such a conclusion must be provisional until these population shifts are analysed in much greater depth.

Chapter Five:

Genescan System Development

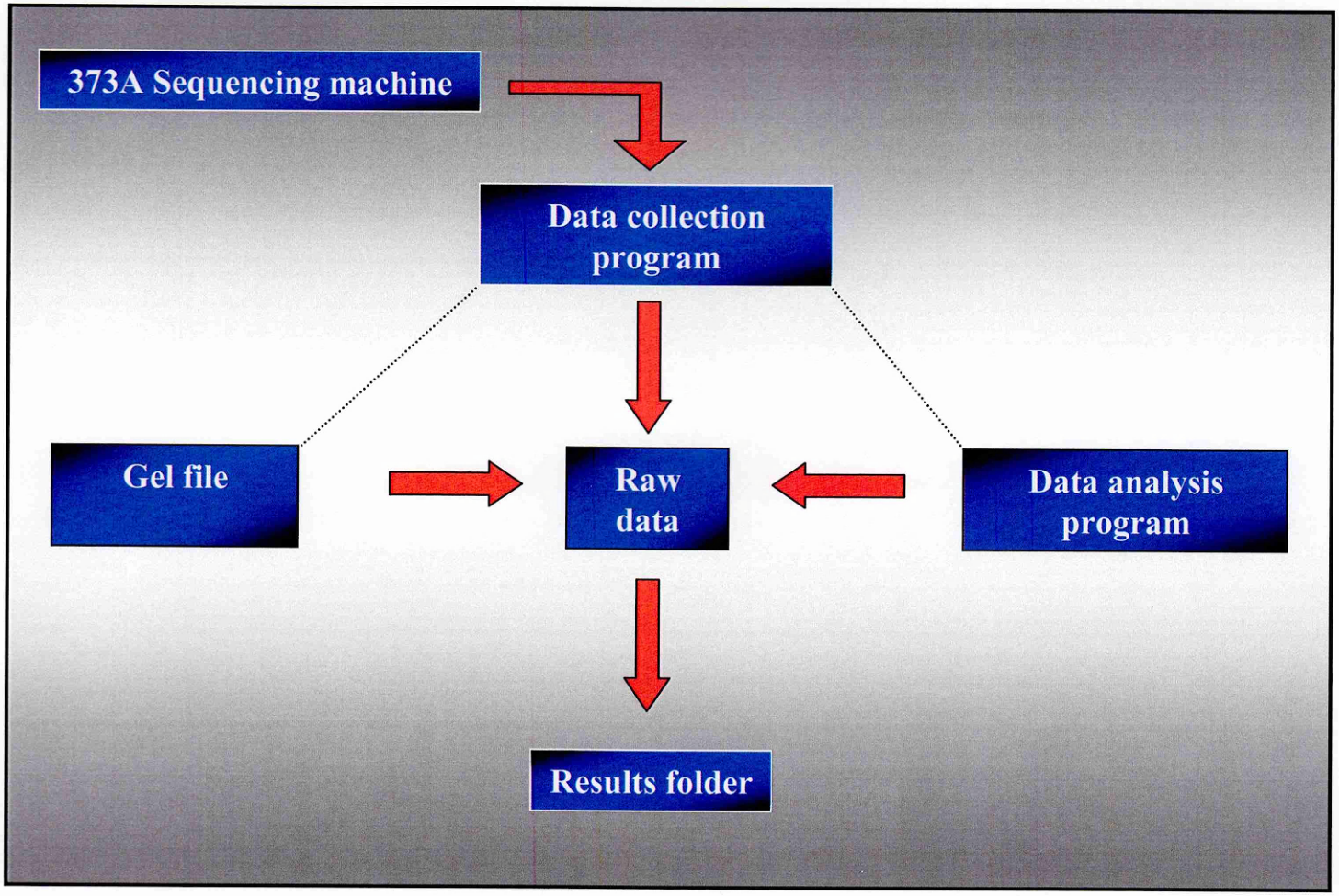
5.1 Introduction

The principal objective of the project is to understand the possible effects of abiotic and biotic stresses on microbial community structure in different soils and on important sub-components of it. To address the molecular aspect of population structure adequately entails screening the PCR fingerprints of a large number of soil bacterial isolates, so that we can be confident that any population shifts observed are significant. To obtain large numbers of accurate DNA fragment length measurements, automation is desirable. The disadvantage with techniques using conventional gels and primers, where PCR fragments are stained with ethidium bromide and visualised under UV light, is that length determination is not very accurate, whether measured by hand or image analysis software. It is known that there is variation in the electrophoretic migration of DNA from one gel to another (variation in the gel or running conditions), or even from lane to lane across a single gel (gel smiling). Another problem is that human interpretation of gels can vary, making comparison between different experiments difficult and problematic. Thus, there are major problems in creating a database with a large number of microbial fingerprints. To overcome these problems, we developed a semi-automated system: Genescan™.

5.2 ABI Genescan™ system

The Genescan™ software is an advanced system which sizes and quantifies DNA fragments by automated fluorescence detection, using laser scanning during acrylamide gel electrophoresis, eliminating the need for radioactivity and post-run gel handling. The software is run on the 373 model of the ABI-Sequencer (Figure 5.1). Fluorescently-labeled PCR samples with differently labeled internal standards are run on the Genescan™ acrylamide gel, results (i.e. product sizes) are automatically supplied in electronic form when electrophoresis is completed. This eliminates the need for data entry and, therefore, the possible introduction of human error.

Figure 5.1 : A conceptual overview of the model 373 software



The system compensates for some of the disadvantages of conventional techniques. It is designed so that each lane on the gel stands alone and is analysed and interpreted separately. The concept of the Genescan™ system is that each lane of the gel can be run independently of all others on the gel, allowing for the independent analysis of each lane. However, the most important aspect is that it allows the analysis of a number of gels which can be cross-compared with confidence. The PCR products are sized electronically by comparison with an internal standard (precision sizing of DNA fragments) in each track; hence no human input is required. The system will allow some manipulation of the data but the raw data can never be altered. These features were originally designed for forensic science and DNA human heredity analysis. However, the system had to be modified to allow us to apply it to the molecular ecology of bacterial communities.

This objective was to modify the system, using the software available, to give reliable microbial DNA fingerprints. In order to apply the system to perform fingerprinting analysis, it was necessary to set it up on the 373 model DNA sequencer. When operational, the next step was to optimize the system parameters to allow development of the PCR genomic fingerprinting technique, in order to create a microbial database of PCR genomic fingerprint patterns of soil isolates.

5.3 Genescan™ parameters

There are a number of different parameters that have to be established to ensure the successful operation of the Genescan on the 373-sequencer machine. All parameters are interactive and therefore must be completed before we can develop this system for our purpose i.e. microbial fingerprint analysis. These parameters are the following;

1. Fluorescent dyes
2. Internal size standard
3. Genescan 2500
4. Gel Matrix

5.3.1 Fluorescent dyes

The Genescan™ system uses fluorescent dyes, which are recognized by the laser in the ABI Sequencer. The fluorescent Dyes are FAM (5-carboxyfluorescein), JOE (2',7'-dimethoxy-4',5'-dichloro-6-carboxfluorescein), TAMRA (6-carboxy-tetramethyl) and ROX(6-carboxy-rhodamine), which are supplied as NHS-esters and 6-FAM (6-carboxyfluorescein) and HEX (6-carboxy-2',4',7', 4, 7-Hexachlorofluoroscein), which are supplied as fluorescent amidite reagents. They are attached onto the oligonucleotide via an aminohexyl linker and are attached to the 5' end of the DNA using a standard DNA synthesis cycle. After cleavage and deprotection, a fluorescent dye-NHS ester is coupled to the primer through the aminohexyl linker (Genescan™ manual). The stability of the labeled primers is increased by protecting them from heat and light, where feasible, during experimental procedures. Up to four different colours can be used for DNA labeling and still allow accurate band matching. This is achieved due to the different absorbency of the dyes and the ability of the laser to distinguish between them. Table 5.1 shows the maximum absorbency of fluorescent dyes used in this study.

Table 5.1 Maximum absorbency of fluorescent dyes in 0.01 M TEAA, pH 7.0

Dyes	Absorbance
FAM	494 nm
6-FAM	495 nm
JOE	527 nm
HEX	537 nm
TAMRA	558 nm
ROX	586 nm

Note: The FAM and JOE dyes have decreasing absorbency as the pH becomes more acidic, so it is important to keep the pH neutral or slightly basic.

Design of fluorescently labelled primers: The principles for designing these labeled primers are the same as for primers used in traditional PCR procedures. The primer sequence should be chosen to maximize the stability and specificity of binding to the desired template location, while avoiding the destabilisation of internal structures. There are three aspects of primer design which should be addressed: binding specificity, base composition and primer internal structures.

1) Base composition

G-C bonds contribute more to the stability (i.e. increased melting temperature) of a primer/template hybridization than do A-T bonds. This is illustrated in the equation below:

$$\mathbf{Tm = 2P + 4Q}$$

Where **Tm**: Melting temperature (°C)

P: Number of A: T pairs

Q: Number of G-C pairs

(Suggs *et al.*, 1981).

2) Primer internal structure

The order of bases will influence the overall primer/template complex stability.

It is desirable to choose primers which contain clusters of Gs and Cs. Also the presence of a G-C at the 3' end of the primer will help stabilise the complex for binding with DNA polymerase. However, as a rule of thumb, runs of more than three consecutive Gs in primers should be avoided. Potential primers should also be examined for self-complementarily and hairpin structures. (Note: a stable hairpin can form with just four G-C base pairs in the stem, and just three bases in the loop).

Once the primers have been designed and the fluorescent label is chosen, then the PCR reaction has to be optimised. The fluorescently labeled primers must be tested so as not to interfere with the efficiency of the reaction. This procedure is reviewed in more detail in Chapter 8.

5.3.2 Internal size standards

The Genescan™ system relies on internal standards in each lane to make it a semi-automated system for analysing PCR products and eliminates the problem of band-shift artifacts and gel-to-gel variation, which are often encountered with other techniques.
The inclusion of an internal lane standard results in three important advantages:

1) It serves as a positive control for the scanner analysis. Even if the PCR fails to amplify properly, normal detection of the internal lane standard indicates proper scanner detection.

2) The internal lane standard is used to size and quantify PCR products. For sizing, the standard is used to create a calibration curve of peak arrival time, which in turn is used to calculate the length of an unknown PCR product automatically. Sizing is accurate and precise. The internal standard effectively controls the lane-to-lane gel electrophoresis mobility variations.

3) The standard can be used to normalise detection of the signal for quantitation. Once the molar amount of internal lane standard added to a sample for scanner analysis is known, relative signal intensities can be used to estimate the molar amounts of an unknown PCR product.

The signal level is important, especially when testing the amount of labeling required for PCR genomic fingerprinting. The Genescan system is sensitive to signal intensities with negative effects resulting from both over and under labeling of PCR products with fluorescent dyes. The level of fluorescent label attached to the PCR products is important and has to be optimised, for routine analysis of labeled products to be performed.

Genescan 2500 Size Markers: The PCR products are end labeled and analysed via fluorescent dyes that are different in colour from that of the internal lane standard. Each labeled PCR product is combined with an aliquot of the internal lane standard and electrophoresed on the ABI sequencer gel. However, as a prerequisite, a selected size standard must first be used to create a calibration curve within each lane.

There is a range of internal size standards available with different lengths but the Genescan 2500 standard is the most relevant size standard for the product sizes that we intend to produce. Genescan 2500 standard is made from lambda DNA digested with *Pst*I followed by ligation of either a TAMRA (yellow) or ROX (red) oligonucleotide to the cut ends. It has 28 fragments ranging from 55 to 14097 base pairs (bp). Fluorescently labeled non-denatured fragments run as though they are 18 nucleotides longer than denatured fragments. The standard molecular lengths are shown on Table 5.2.

Experimentally, the gel type found to be most suitable for routinely sizing microbial fingerprint PCR products was a 4.75% acrylamide gel, run under non-denaturing conditions (section 5.5.1). The non-denaturing gel was chosen because it had better resolution for larger size PCR products as the band sizes for genomic fingerprinting can range between 50bp to 4.0 kp. The size calibration curve for Genescan 2500 was produced for the automatic sizing of products. However, gel variation may cause problems with the calibration curve, so it is important to check the size standards in each run. Figure 5.2 shows a computer-generated image of Genescan 2500 run under non-denaturing conditions using a 4.75% acrylamide gel.

Figure 5.2 A computer-generated image of Genescan 2500, run under non-denaturing conditions using a 4.75% acrylamide gel recorded by Genescan™ with ERIC Profiles of soil isolates

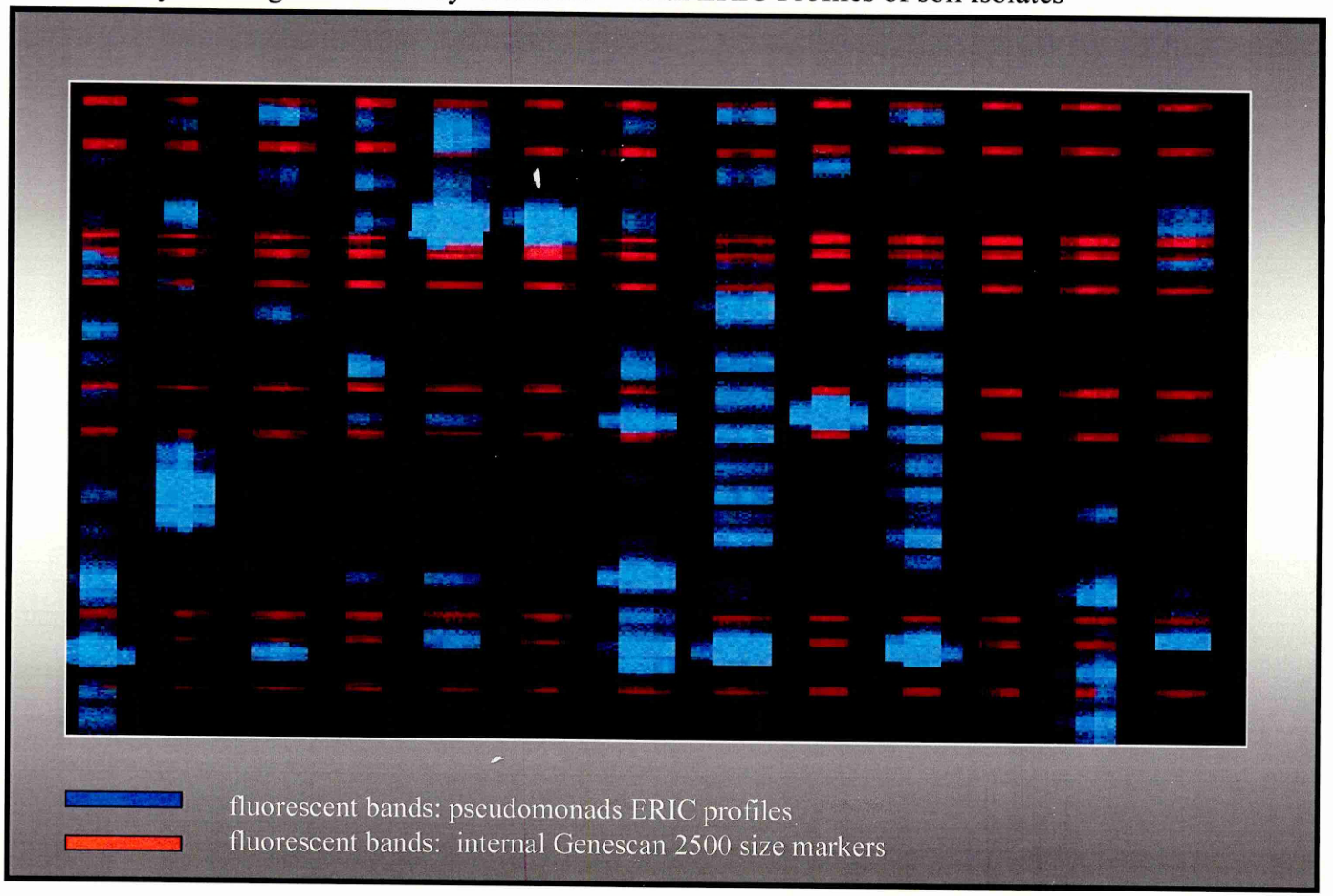


Table 5.2 Genescan 2500 standard molecular lengths (bp) for running on denatured gels and non-denatured (native) gels

Denatured	Non-denatured
14079	14097
5099	5117
4771	4789
4529	4547
2860	2878
2481	2499
2465	2483
2162	2180
2008	2026
1722	1740
1181	1199
1115	1133
827	845
536	554
490	508
470	488
361	379
286	304
269	287
238	256
233	251
222	240
186	204
172	190
116	134
109	127
94	112
37	55

5.3.3 Matrix file

There are subtle differences in the way Genescan™ and the sequencing software collect and interpret data. Therefore, the machine must be calibrated to recognise the dyes, gel type and the sequencing machine used. This type of calibration is known as the Matrix file. A new matrix must be made for each dye set, gel type, gel conditions and filter colour combination used with the Genescan™ software. A specific Genescan™ gel must be set up so that a matrix file is created for analysing fragment data collected during subsequent routine running of Genescan™ runs.

Creation of a matrix file: a matrix file is created only when the running conditions for the gel and the sequencing machine have been standardised (section 5.5). Once these parameters are obtained, then four fluorescent dyes must be chosen (the 373 system has a filter wheel system which restricts the number of dyes to four). The Dye Primer Matrix Standard and the Fluorescent Amidite Matrix Standards are commercially available and are used to calibrate the machine for both the dye colour and the fluorescence of each dye (Dye Primer Matrix Standard Kit P/N 401114, Fluorescent Amidite Matrix Standards, P/N 401456). It is important to ensure that the appropriate filter set is selected for the dyes which are to be detected. An aliquot of each dye is loaded onto separate lanes on a standard gel and run using standard running conditions.

The data is collected via the GENESCAN (GS) collection program and analysed using the GS Analysis program. It is important that the auto Lane Tracking and Baseline Data options are selected and the multi-Component option is not selected. This allows for the calibration of the software.

Each lane is set and tracked, and the peak height of each dye colour recorded by the matching filter which matches the dye colour (the highest peak height taken for the chosen dye is assumed to be when all other dyes give the lowest peak readings). Once analysed for each of the dyes, a matrix file is created for the gel type. In summary, the Genescan™

software is an advanced system. Once the parameters have been set, this semi-automated system will allow for the rapid throughput of data from the microbial isolates.

5.4 Methods

5.4.1 Genescan™ acrylamide gel preparation

The ABI-Applied Biosystems Systems 373 Sequencing Machine is a real-time electrophoresis detector. A laser scans across a region of the gel during the electrophoresis process, detecting the fluorescently labeled DNA passing that region. The system scans the gel width 600 times per hour and when a labeled DNA fragment migrates into the laser-scanning region, a photomultiplier tube (PMT) detects fluorescent light and converts it into an electrical signal. These signals are then transmitted to the computer and stored for future processing.

Because the laser must pass through the gel, the glass and the gel must therefore be non-fluorescent. Extra care must be taken to eliminate any particulate or fluorescent matter from the acrylamide gel solutions and the glass plates must be kept extremely clean and free of dust. It is important always to use ultra-pure reagents. The ABI-Applied Biosystems System recommends all reagents be supplied by BIOMETRA™. High-grade, distilled deionised water is used to prepare all solutions, which are filtered to remove any particulate matter that may fluoresce or scatter light.

Abbreviations:

dH_2O	deionised, distilled water
TBE	Tris-borate-EDTA (pH 8.3)
TEMED	N, N, N', N'-tretramethethylethylenediamine
DTT	dithiothreitol
TEAA	triethylammonium acetate
dNTP	deoxynucleoside triphate
ddNTP	dideoxynucleoside triphosphate

5.4.2 Reagents and solutions

1) 10X TBE stock solution (per litre)

Tris Base	108.0g
Boric Acid	55.0g
Na_2 EDTA	8.3g

Working solution (1X) is 89 mM Tris, 89 mM Boric acid, 2 mM EDTA; pH should be 8.3 at room temperature.

2) Acrylamide stock solution 40% (19:1)

Acrylamide	38.0g
Bis-acrylamide	2.0g

3) 4.75 % non-denatured gel mix

40% acrylamide Stock Solution	7.2 ml
dH_2O	43.4ml
Mixed bed ion-exchange resin	1.0g

5.4.3 Gel preparation and casting equipment

The glass plates, gel spacers and comb are washed with warm water, taking care to remove all traces of detergent. The equipment is thoroughly rinsed with dH_2O and left to air dry. The plates are placed on an elevated box, the spacers arranged and the corner notch section taped to ensure that the ears of the gel are formed correctly.

The 4.75 % non-denatured gel mix is stirred at room temperature for 1 hour, 6 ml of 1X TBE is filtered through a 0.2 um filter unit. It is important to add the acrylamide mix after the TBE is filtered, otherwise this seriously affects the quality of the gel. The acrylamide gel mix is filtered through a 0.2 μm cellulose nitrate filter for approx. 10 min to degas.

5.4.4 Gel casting

When the gel mixture is degassed, 300 µl of freshly made 10% ammonium persulfate solution is added and swirled gently to mix, avoiding air bubble formation. Then 33µl of TEMED is added and swirled gently. These solutions start the polymerization of the gel, so the gel must then be cast immediately. When pouring acrylamide gels, it is important not to allow air bubbles to form as they have the following effects on an acrylamide sequencing gel; The efficiency of gel setting can cause interference with the running of the fragments in the gel; and problems with the laser's ability to read the gel (i.e., if the air bubble occurs in the 'read region' where the laser scans during the run).

The gel is poured slowly, but at a constant rate using a 60 ml syringe containing the gel mixture. Tapping the plates enables the trapped bubbles to rise to the surface of the gel and thus avoid air bubble formation.

The level of gel solution must fill the plates, thus allowing the current to flow through the gel. The GENESCAN 36-tooth comb is inserted carefully to avoid introducing any air bubbles as they affect the well formation. Then the gel is left to set. A minimum of two hours is needed to ensure complete polymerisation of the gel prior to running it. The gel can be stored for a maximum of 24 hours before the resolution of the gel deteriorates, but fresh gels are recommended.

5.5 Results

5.5.1 Genescan™ gel

As mentioned earlier, the running conditions of the gel were the first parameter to be set. A number of different percentage acrylamide gels were tested varying from 6% to 4% acrylamide, both denatured and non-denatured. It was found that the 4.75% non-denatured gel was the most satisfactory, because it gave good resolution for the PCR products (genomic fingerprinting range between 50bp to 4.0kp). The following conditions were chosen as standard for the Genescan gels for microbial PCR fingerprinting.

Running conditions

Gel Type	4.75 % non-denaturing acrylamide gel
Volts	2500
Current	45mA
Power	30 Watts
PMT	512
Laser	40 mn
Scan read	from 0 to 10,000
Running time	10 hours
Gel thickness	4mm
Read distance	32 cm

5.5.2 Matrix file

The matrix was created when the running conditions were set. The matrix was set for the following fluorescent dyes: Tamara (yellow), Rox (red), 6-Fam (blue) and Joe (green). The values were chosen specifically so that the maximum reading was achieved for one dye when the readings for the other three dyes are at there lowest. Thus decreasing the amount of interference between the difference dyes. The following values were set and calibrated for these dyes (Table 5.3).

Table 5.3 Values used for calibration of fluorescent dyes

Matrix values for a 4.75% native acrylamide gel				
	Blue	**Green**	**Yellow**	**Red**
Blue	1.00	0.2634	0.0081	0.001
Green	0.5455	1.00	0.2991	0.0082
Yellow	0.3017	0.5659	1.00	0.1928
Red	0.2470	0.2470	0.4945	1.00

This matrix file was to reflect these specific running conditions. The matrix file was then connected to the 373 machine to enable automatic recognition of these dyes. However changing the running conditions would affect recognition of the dyes, requiring a new matrix to be created.

5.5.3 Internal size standard

With the standard gels a calibration curve for the internal size standards was created. The calibration curve was set using the Genescan 2500 (see table 5.2). This is important, as the length of each PCR product is determined by comparison with the calibration curve in the specific lane in which it was run. A number of methods are available to calculate the size of the PCR products. Two methods recommended for non-denatured gels are the "Southern Global Method" and the "Southern Local Method" (Genescan 672 Software User's Manual).

5.5.3.1 Southern Global Method

This creates a best-fit line through all the points and then uses values found on that line to calculate the fragment values and it compensates for the standard fragments which may run anomalously.

The equation used for calculating the **Southern Global Method** is:

$$L = [c/(m-m_o)] + L_0$$

Where m =mobility

L_o= the length of the standard fragment

The Lo and mo and c are calculated by a least squares fit to minimize the quantity,

$$\Sigma_i(L_i-(c/(m_i-m_o)+L_o))^2$$

(Ref: Genescan manual)

5.5.3.2 Southern Local Method

This Southern Local Method uses the three standard fragments closest in size to the unknown fragment to determine a best-fit line value. However, if any of the standard fragments run anomalously, size estimates based on this method may be inaccurate.

The equation for calculating the **Southern Local Method** is:

$$L = [c/(m-m_0)] + L_0$$

Where M =mobility

L_0= the length of the standard fragment

(Ref: Genescan manual)

The Southern Local Method is recommended for the routine analysis of PCR fingerprints. However, if any of the PCR products interfere with the standard size fragments, then re-analysis with the Southern Global Method can be employed. It is usually recommended that the size standards be checked on each gel to decide if the run conditions are acceptable.

5.6 Discussion

The methods employed by Genescan™ for the sizing of PCR products are highly accurate and reproducible. The automatic size determination obtained with the use of an internal lane standard demonstrates one advantage of this multi-colour fluorescence detection. Once a sample is loaded onto the gel, no further manipulation of either gel or data is necessary in order to determine the size of unknown DNA fragments; the result is then obtained as a permanent computer or hard-copy record. Hence the results have no operator bias, allowing the construction of a database of microbial fingerprints which can be compared with more confidence.

5.7 Conclusions

The Genescan™ system was successfully set up on the 373 ABI Sequencer. In this chapter, I have shown how the basic experimental work was completed and system parameters set up. Standard technique now ready to be employed by different users and for different applications. The completion of this section enabled the study to proceed with development of the PCR fingerprinting of soil isolates, and the application of the Genescan™ automated system to these fingerprints, as we shall go on to demonstrate in Chapter 8.

Chapter Six:
Access Database Development

6.1 Introduction

This next stage of the project required the development of an analysis method. It had to be capable of analysing the Genescan™ data files, and the Microsoft™ Access computer program was chosen for this process (Table 6.1). To create a database capable of manipulating the data into a user-friendly format, I devised the 'Genescan™ Microsoft database'. It was designed to mimic some aspects of the Perkin-Elmer Applied Biosystems Genotyper™, commercial software intended to analyse Genescan™ data files specifically for forensic science and DNA human heredity. Thus, the specifications were not entirely suitable for our application since one of the primary process objectives of this project was to create a means of producing a database of microbial fingerprints. This required turning complex data into a simple binary data format (i.e. a series of 1 and 0, one if band is present and zero if not), which can be then used by all other computer programs.

6.1.1 Definition

A database is a collection of information related to a particular subject or purpose. Information can be managed from a single database file as follows:

1) Data is divided into separate storage containers called **tables**
2) The table data is then manipulated using **queries**; to view, change, and analyse data in different ways
3) The results from the queries are then analysed and printed to a **report file**

It is designed to present data in a certain way (as indicated in the final query), the final report can then be exported out of Microsoft Access into other computer packages.

6.2 Principles

Good database *design* is the keystone to creating a database which performs effectively, accurately, and efficiently and needs expert advice. The database developed for this

project was designed in conjunction with the Modeling Group of the Soil Science Dept, IACR-Rothamsted (O'Flaherty and Coleman 1999).

6.3 Design steps

There are seven basic steps in designing a database:

1. Determining the purpose of the database
2. Determining the tables required
3. Determining the fields needed
4. Identifying the fields with unique values
5. Determining the relationships between tables
6. Refining the design
7. Adding data and creating the database

6.3.1 Purpose

The first step in designing a Microsoft Access database is to determine the purpose of the database and how it will be used. I required the database to take the complex data (Table 6.1) and turn it into a simple data form (binary format). Data processed in this way can then be transferred to other computer program packages for further analysis.

Table 6.1 Genescan™ data files showing the computer data provided by the Genescan system.

Sample	d1	Base	Peak Height	Peak Area	d5
fb1.9	110.9	128.04	1227	12157	1109
fb1.9	112.9	133.23	296	2501	1129
fb1.9	113.5	134.8	372	2474	1135
fb1.9	114.1	136.4	400	4752	1141
fb1.9	116.9	143.9	346	2425	1169
fb1.9	117.5	145.52	390	4155	1175
fb1.9	121.9	157.42	121	1007	1219
fb1.9	122.5	159.06	94	814	1225
fb1.9	125.3	166.71	103	578	1253
fb1.9	125.9	168.36	138	878	1259
fb1.9	134.7	192.7	63	391	1347
fb1.9	135.5	194.87	96	973	1355
fb1.9	141.3	211.68	91	467	1413
fb1.9	141.9	213.54	131	822	1419
fb1.9	143.3	217.9	373	6718	1433
fb1.9	150.4	240.61	57	719	1504
fb1.9	152.7	247.63	539	10475	1527
fb1.9	156.4	259.17	645	3554	1564
fb1.9	157.1	261.38	842	14348	1571

fb1.9	161.1	274.14	210	1567	1611
fb1.9	161.7	276.06	220	3378	1617
fb1.9	163.3	281.2	205	1633	1633
fb1.9	164.3	284.42	246	1908	1643
fb1.9	164.8	286.03	340	1922	1648

Therefore it was necessary to formulate a series of tables and queries to achieve this objective.

6.3.2 Tables

Determining the tables is a difficult step in the design process. The results wanted from the database, i.e. the report that will be exported, and the questions that will be answered, do not necessarily provide clues about the structure of the tables which produce them. A table should not contain duplicate information, and information should not be duplicated between tables. Therefore, when updating information it should be only stored in a single table. This makes it more efficient, and also eliminates the possibility of duplicate entries that might contain different information.

6.3.3 Fields

Each table contains information about the same subject, and each field in a table contains individual facts about the table's subject. Four critical steps are to:

· Ensure each field relates directly to the subject of the table

· Avoid including derived or calculated data (data that is the result of an expression).

· Include all the information that is needed.

· Store information in its smallest logical parts.

6.3.4 Fields with unique values

In order for Microsoft Access to connect information stored in separate tables, each table in the database must include a field, or set of fields, which uniquely identifies each individual record in the table. That means that once the different tables for each subject in the database are set up, there must be a way of telling Microsoft Access how to bring that information back together again. The first step in this process is to define relationships between the tables. Once completed, queries, forms, and reports can be created to display information from several tables at once.

6.3.5 Relationships between tables

Once the information is divided into tables, and the primary key fields are identified, Microsoft Access must be instructed how to bring related information back together again in meaningful ways. To do this, relationships between tables via queries must be defined.

6.3.6 Refinement

Once the tables, fields, and relationships are designed, then the criteria set up must be tested, to detect any flaws and ensure that the results are in the expected format.

6.3.7 Data entry & creation

Once the above structures meet the design goals, then I can submit existing data to the tables and create a database.

6.4 Results

The microbial fingerprint database was set up to contain a number of interacting tables and queries. Each table interacts with each query until the final query is reached and the data has been manipulated and analysed. The database is programmed to turn the base pair sizes obtained in Genescan into a series of 1 and zero. The value of 1 is given if a PCR product is within the select segment, and zero if there is no product (Table 6.2).

The first table accepts data files from the ABI Genescan™. The next four tables create lists of the different factors (list of data, groups of base pairs as decided from the level of fluorescence from the gel files). Next it lists the isolates, and finally a summary of the data, i.e. isolate name, bases pair segments (the block size is set at 10 bp). More importantly, it records a 1 or 0, depending on the presence or absence of a band in the blocks of base pairs (as defined in the queries).

Table 6.2 Genescan file analysed by the database program, produced the following result

Isolate	Segment	Presence
fb1.9	100	0
fb1.9	120	1
fb1.9	140	1
fb1.9	160	0
fb1.9	180	0
fb1.9	200	1
fb1.9	220	0
fb1.9	240	1
fb1.9	260	1
fb1.9	280	1
fb1.9	300	0
fb1.9	320	0
fb1.9	340	1
fb1.9	360	0
fb1.9	380	0
fb1.9	400	0
fb1.9	420	0
fb1.9	440	0
fb1.9	460	1
fb1.9	480	0

Each query is required to reset the data in each of the tables. This is where the mathematical element in the database system is employed. The first query is important and it requires the following to be set:

(1) What range of size products should be read? (e.g. between 100 and 1500 bp)

(2) What level of peak height is acceptable? This sets at what point I accept a peak to represent a PCR product and also the noise level, which will remove all backward and forward stutters from the Genescan™ gel. This query requires reset after each use, depending on the quality of labeled PCR products.

(3) At what intervals to set the base pair segments? The range can be divided into 128 segments, where each segment represents 10bp. It is worth noting that: 10 base pair differences would not be visible on an agarose gel.

Then the next four queries instruct the creation of the groups, list of isolates and presence and absence of bands for each isolate, and the crossbar table of results from the analysis. The final query instructs the creation of the result report which will produce the final analysed data in the requested format.

The report is a way to present data in a printed format which can be exported to other computer applications (Table 6.2). The data above shows a list of soil isolates with the base pair segments. This preliminary design of the database analysis system requires a number of adjustments during development of the microbial fingerprint database. The data is in a binary format, which can now be transferred to another software package for the next step in the analysis of the fingerprint data.

6.4.1 Standardised analysis parameters

The Genescan Fingerprint Access Database (O'Flaherty and Coleman 1999) was successfully used for all analysis of the fingerprint data (which is reported in Chapter 9). The binary format allowed the data to be successfully analysed by Genstat, ERIC and PHYLIP programs. Standardised analysis parameters, determined by analysing data of various parameters, were set as follows:

- size of product range set between 100 and 1500 bp
- the level of peak height (noise level) set at 150
- intervals set at 75 bp segment

The criteria for accepting the parameters must be reviewed for each experiment and cannot be universally imposed. Due to the nature of the analysis packages used (Genstat, ERIC and PHYLIP), parameters have to be standardised to allow for direct comparison between all the fingerprint data (i.e. all different time points).

6.5 Conclusions

The conversion of automated computerised data obtained from the ABI GENESCAN™ system into a format capable of being used by the Phylogenetic Programs is the first, critical step. The data, subsequently converted to binary format by the database, allows us to use these Phylogenetic and IBM-compatible programs for analysis. Furthermore, the database is designed to allow us to carry out the initial stages of analysis, giving us the ability to determine data criteria which will be acceptable across a range of phylogenetic programs.

Chapter Seven:

Fingerprint data analysis

7.1 Introduction

The next step in the development of the analysis strategy was to infer a phylogenetic relationship between isolates fingerprints. To do this successfully, it was necessary to develop a computer program which makes statistical comparison between microbial fingerprints using the binary data. The next stage of the project was writing of Genstat programs to analyse the data statistically. These programs were written in conjunction with the Genstat programmers at IACR-Rothamsted, and a range of programs have been written. The programs create distance matrixes which can be read by the phylogenetic program Phylip™ (Felstein, 1994), and which then are converted to dendrograms. These are generally used in the literature to infer a phylogenetic relationship between groups of organisms. The method used to infer relationships is the 'distance matrix method'. This method is algorithmic and refers to dissimilarity measurements between organisms (this mathematically relates the organisms to each other).

There is a long-running debate about which mathematical equation, from a number of different equations available to create a matrix to compare data (Priest and Austin, 1993), should be used to create a distance matrix when analysing data. However, the most common in the published material reviewed for this project were Jaccard and Dice equations.

Table 7.1 Distance matrices formulas used to confer relationships between microorganisms

Jaccard	a /a+b+c
Dice	2a/ 2a+b+c

Note: Both equations discount negative pairing, but Dice gives double weighting to positive pairing.

a: corresponds to the number of positive matches.

b and c: represent the number of non-matching characters between operational taxonomic units.

The main difference between the two equations is that Dice is written so that it gives extra weighting to a positive band match.

			weighting				weighting
Jaccard	if ai = aj = 1,	then 1	1	**Dice**	if ai = aj = 1,	then 1	2
	if bi = bj = 0,	then 0	0		if bi = bj = 0,	then 0	0
	if ci = cj,	then 0	1		if ci = cj,	then 0	1

(Genstat 5 Release 4.1 (Third Edition))

7.2 Methods

Initially, we programmed for Jaccard and produced a **Dissimilarity Matrix Program**, (1- Jaccard Equation). Initial testing of these programs made us consider the type of data that we were analysing; we found that the Dice equation was more applicable to the data being produced. In simple terms, we are more interested in positive band matching than non-matching bands and the Dice equation gives extra weighting to positive matches. (If we were also interested in the non-matching bands, we would have employed the Jaccard equation as well). Therefore the program has been modified to allow the data to be analysed using the Dice co-efficient, instead of Jaccard. This program is simply called the **'Dice Program'**, with an option for a similarity or dissimilarity matrix.

A program has been written to calculate statistically the confidence interval of each cluster in the phylogenetic trees, produced by the phylogenetic programs. This formula was recommended by Grothues and Tümmler (Grothues *et al.,* 1991) and has been modified by J. Fenlon (personal communication, 1998). This program is called **Confident Interval Matrix Program** (Appendix A; Grothues *et al.,* 1991). We have a program called **Dendrogram Program**, which allows us to draw phylogenetic trees in Genstat with a calibration index (Appendix A). With these systems in operation, preliminary analysis of the data was possible.

7.3 Results

7.3.1 Evaluating the computer programs

Having established a means of creating distance matrixes which can be read by the phylogenetic program Phylip™, it was necessary to examine a number of different methods used to infer phylogenetic relationships in groups of organisms. These programs were compared to find a way of conferring a meaningful phylogenetic relationship between microbial fingerprints. A phylogenetic tree shows the evolution or the ancestry of organisms, expressed via cluster analysis and dendrograms. *Cluster* refers to the mathematical determination of groups from a distance or similarity matrix. The *dendrogram* is a hierarchical tree, which expresses relationships between clusters. The Genstat™ similarity and dissimilarity distance matrixes were used to construct a dendrogram using the PHYLIP software package Version 3.1 (Felstein, 1996).

It is difficult to assess the effect of the different programs on experimental data, as we do not know the relationships between isolates and have no means of checking the relationship implied by the phylogenetic trees. Therefore, to gain an understanding of the Phylip Program Suite, and also to study the way it confers phylogenetic relationship between isolates, test data was created to reflect the type of data that is obtained from microbial fingerprints. The data was designed in a very simple system of band patterns to demonstrate the relationship between the fingerprints and help us to understand the clustering programs.

Table 7.2 Test RAPD data used to test phylogenetic relationships between fingerprints of isolates. (Test 1 is not profiled, because it has a positive band in every site from 100 to 15000, i.e. a band in every 10 base segment).

Sample	Base number (bp)				
	100	400	600	800	1200
Test 2	1	1	1	1	1
Test 3	**1**	**1**	**1**	**1**	**0**
Test3a	**1**	**1**	**1**	**1**	**0**
Test 4	**1**	**1**	**1**	**0**	**0**
Test 5	**1**	**1**	**0**	**0**	**0**
Test 6	*0*	*1*	*1*	*1*	*1*
Test 7	*0*	*0*	*1*	*1*	*1*
Test 8	*0*	*0*	*0*	*1*	*1*
Test 9	*1*	*1*	*0*	*1*	*1*
Test 10	*1*	*0*	*1*	*1*	*1*

7.3.2 Data description

Test 1 is designed to be an "out-group", which means it should show no relationship to the rest of the data set. Test 2 is set to test the clustering programs, as it is difficult to group. It has common bands to the other isolates and therefore may go into more than one cluster. Likewise, test 9 is also difficult to match (it is missing the 600 bp band). This makes it a target for both groups, but its nearest neighbour should be isolate 2 as they have only one band different. The rest of the data has been designed so that the top group are similar to each other, and hence can be grouped in that order (**in bold**) and the bottom section should group together (*in italics*). Here, in theory, is the phylogenetic tree that we would predict:

Cluster 1 Test 1 (Outgroup)

Cluster 2 Test 3,3a, 4,5.

Cluster 3 Test 6,7,8.

Test 2 and test 9 should be clustered together between cluster group 2 and group 3.
Test 10 should cluster near test 6 and 7.

7.4 Phylogenetic programs in PHYLIP

The binary data was analysed in Genstat and a Dice dissimilarly matrix was created. The Dice dissimilarly matrix was imported into Phylip, where there are ranges of phylogenetic packages, to give different phylogenetic trees. We investigated three phylogenetic programs that had been suggested in the literature. A brief description of each program is given, and the phylogenetic trees obtained are then explained.

7.4.1 Neighbour program

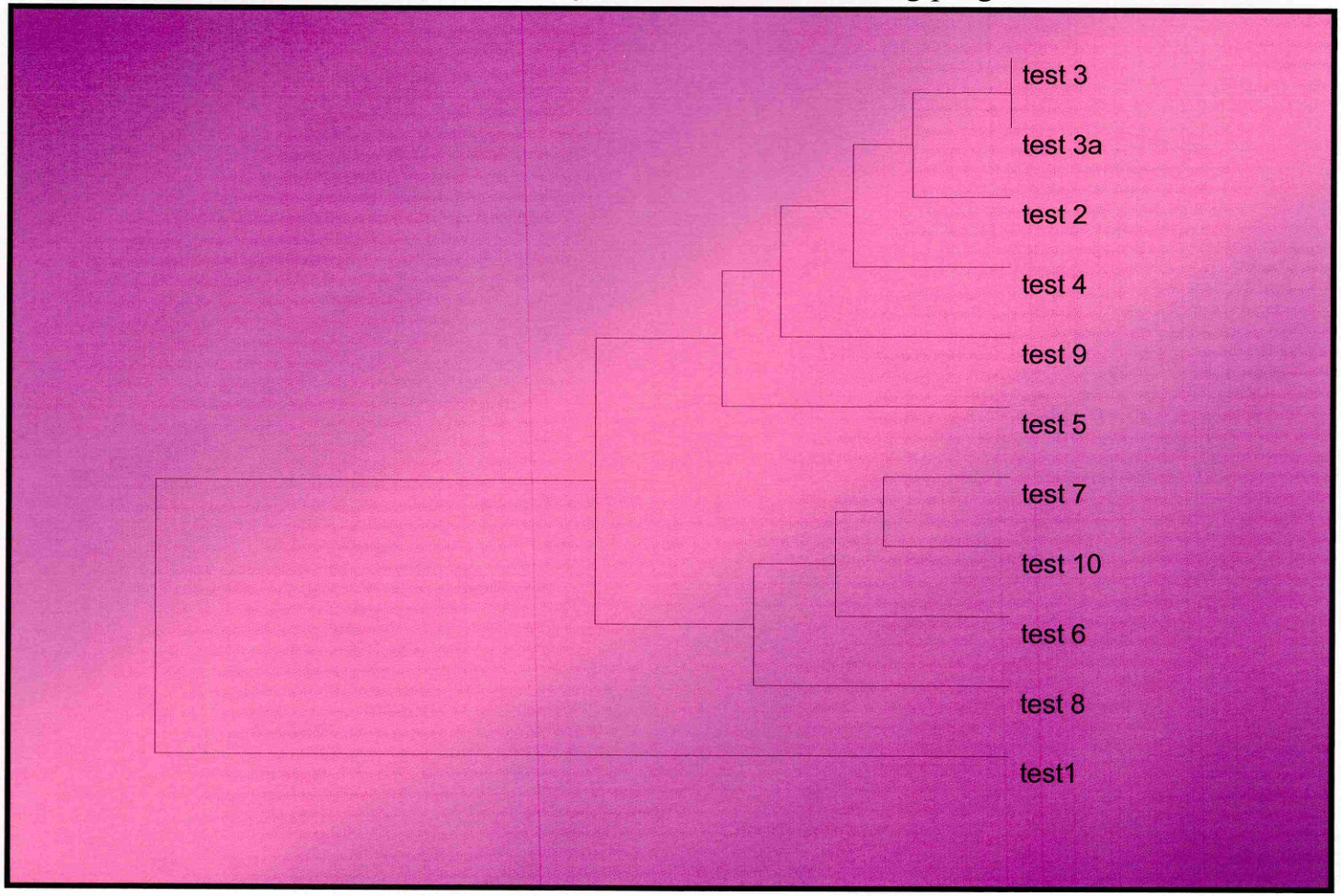
The Neighbour program has two options within it, namely the Neighbour Joining Method (Nei 1987), and UPGMA Method (**U**nweighted **P**air-**G**roup **M**ethod using arithmetic **A**verages, Sneath and Sokal 1973).

1. Nei's "Neighbour Joining Method" constructs a dendogram by successive clustering of lineages, setting branch lengths as the lineages join, and it is not rearranged thereafter. The dendogram does not assume an evolutionary clock, so it is in effect an unrooted tree (Felstein, 1996).

2. UPGMA method constructs a tree by successive (agglomerative) clustering, using an average-linkage method of clustering. UPGMA does assume an evolutionary clock. The branch lengths are not optimised by the *least square* criterion, but the methods are very fast and thus can handle much larger data sets (Felstein, 1994). UPGMA has the ability to handle large amounts of data and is a very fast method.

The neighbour program, with the UPGMA method option, produced a phylogenetic tree which divided the data into three clusters. Figure 7.1 shows the phylogenetic tree produced by the **UPGMA** clustering program.

Figure 7.1 Phylogenetic tree produced by the **UPGMA** clustering program



Output:

Cluster 1 1

Cluster 2 5, 9, 4, 2, 3 and 3a

Cluster 3 8, 6,7, and 10

We obtained the expected three clusters. Cluster 1 and Cluster 3 are as expected (Group 1 Test 1 has been out-grouped. Group 3 test 6,7,8,10 have been placed together). However, we were not as satisfied with Cluster 2. The positioning of test 2 and test 9 was unsatisfactory, as was the positioning of test 4 from test 5, which have been separated in the tree. Overall, this method produced a satisfactory tree with the main groupings correct, but rather more dubious sub-groupings.

7.4.2 Kitsch program

This estimates phylogenetic relationships from distance matrix data under the "ultrametric" model, which is the same as the additive tree model, except that an evolutionary clock is assumed. The Fitch-Margoliash criterion and other least squares criteria are assumed. This means that branches of the tree cannot be of arbitrary length, but are constrained, so that the total length from the root of the tree to any species is the same (Felstein, 1996).

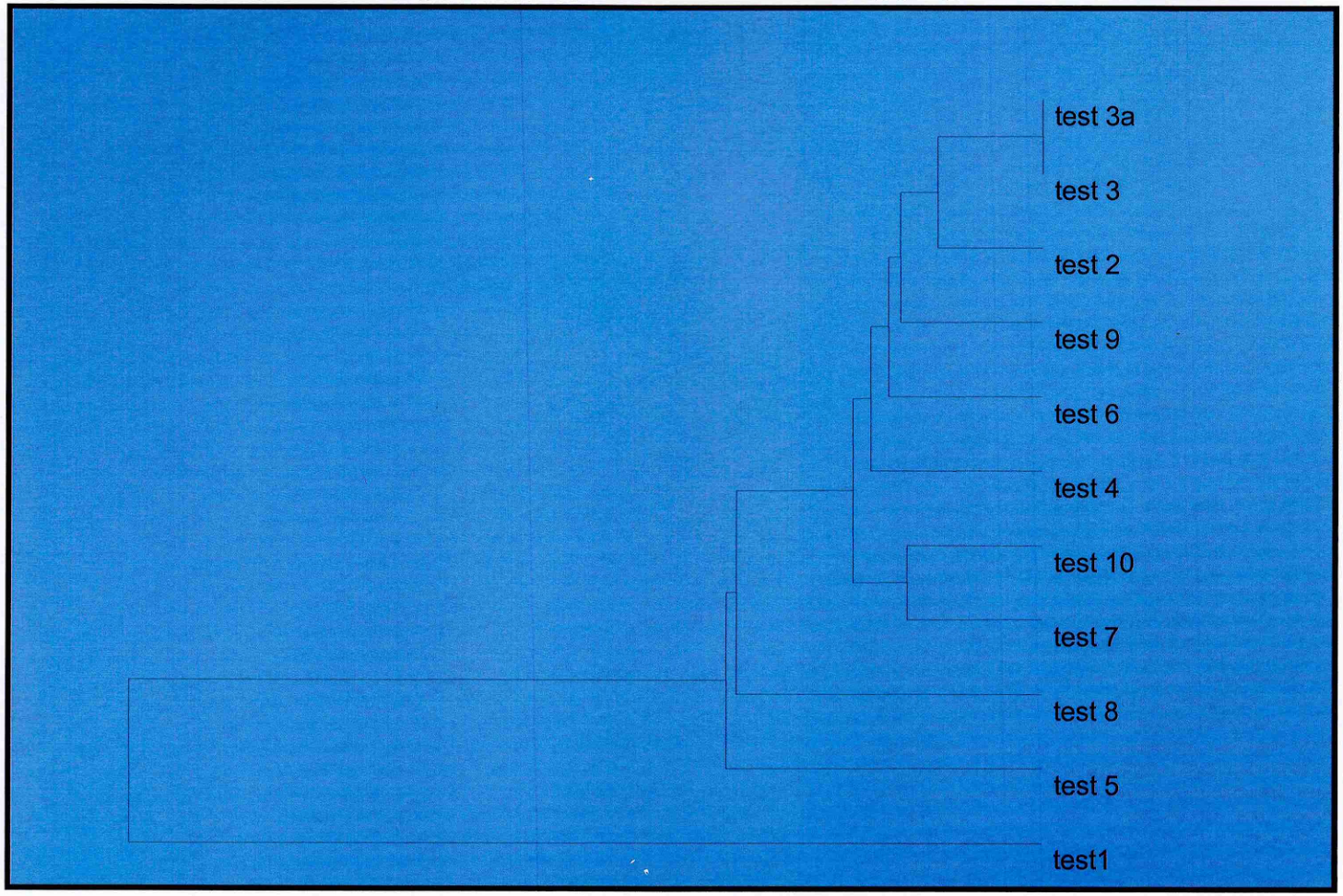
Kitsch method divides the tree into two main clusters. Figure 7.2 shows the phylogenetic tree produced by the **Kitsch** clustering program.

Output:

Cluster 1 1

Cluster 2 5, 9, 4, 2, 3, 3a, 8, 6, 7, 10

Figure 7.2 Phylogenetic tree produced by the **Kitsch** clustering program



In cluster 2 there seems to be three sub-groups:

Sub-group 2.1 3, 3a, 2,9,6,4,10,7

Within this group, test 10 and 7 grouped separately from the others. 3 and 3a are identical. The rest of the isolates are laddered to these identical isolates.

Sub-group 2.2 8

Sub-group 2.3 5

This program is unable to use the data set supplied. Hence it is difficult to extrapolate meaningful interpretations. It was therefore decided that we would not employ this program in any other studies.

7.4.3 Fitch program

This estimates phylogenetic relationships from distance matrix data under the "additive tree model", according to which the distances are expected to equal the sums of branch lengths between species. It uses the Fitch-Margoliash criterion and some related least square criteria, but does not assume an evolutionary clock (Felstein, 1996).

The Fitch method divided the tree into three clusters. Figure 7.3 shows the phylogenetic tree produced by the Fitch clustering program.

Output:

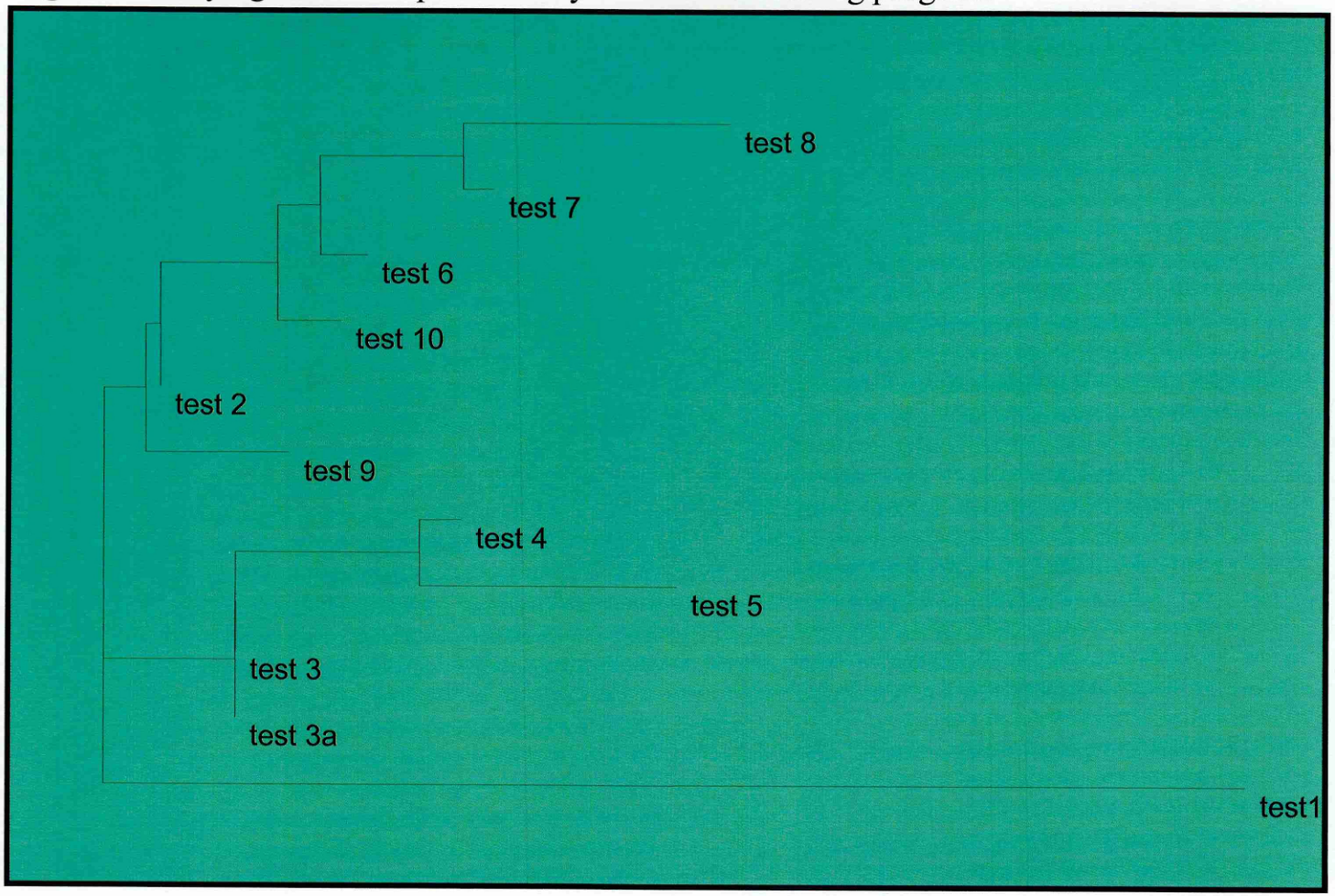
Cluster 1 **1.**

Cluster 2 **5, 4, 3 and 3a**

Cluster 3 **8, 6,7, 9, 2 and 10**

3 and 3a are identical and are related to test 5 and test 4, which are closely related.
Test 8,and 7 are similar, related to test 6 and test 10. Both test 2 and test 9 are in the group but are not strongly associated to the others, which is expected by the position of their bands. Test 2 and test 9 were chosen to confuse the clustering, as they have similarities to the other test isolates.

Figure 7.3 Phylogenetic tree produced by the **Fitch** clustering program



7.5 Discussion

There are concerns over the positioning of the test isolates with the Kitsch program; therefore it is not a suitable program to study the relationship between microbial fingerprints. However, Fitch and Neighbour: UPGMA (Felstein, 1996) produced phylogenetic trees which explained the relationship between the test isolates. The Fitch program has stronger resolution power to characterise the relationship between isolates. While the UPGMA produced a phylogenetic tree with the correct cluster, we were sceptical of some sub-groupings produced in the tree. The Fitch program was excellent for producing phylogenetic trees from our data sets, while the trees produced by the UPGMA program were acceptable.

However, the decision to use routinely UPGMA in preference to Fitch was due to the time taken by the Fitch program to produce our phylogenetic trees. We encountered a problem with the physical running of large numbers in that program, which takes over 7 days and may be terminated in the central computer system due to CPU usage. Investigation into the CPU usage pointed to a program fault in PHYLIP resulting from the complex nature of the data being analysed (J. Felstein, personal communication, 1999). Therefore the overall advantage of the UPGMA is the time taken to produce trees and the fact that it prevents the occurrence of program faults. However, if there is a need to study a sub-population or group it may be feasible to use the Fitch program; for routine analysis, however, we employ the Neighbor-Joining program UPGMA.

7.6 Conclusions

A Microsoft Access database was successfully built, taking complex data from the Genescan™ system and turning it into simple binary data. This was followed by the development of statistical programs in Genstat, allowing us to establish relationships between different isolate fingerprints. Finally, the phylogenetic programs have been tested to determine which is most suitable for routine analysis of fingerprint data. This provides the basis for the project to move to the next stage, studying the abiotic stress of heavy metal contamination in agricultural soil.

Chapter Eight:
Fingerprint Method Development

8.1 Introduction

This chapter reviews the development of methods which have been used to assess the genetic fingerprints of the bacterial population selected in Chapter 3: the pseudomonad group. The technique is **PCR Fingerprinting**, whereby sections of DNA are amplified to generate a fingerprint unique to each bacterial isolate (section 1.2.3). A rapid and reproducible method for genomic fingerprinting is the repetitive extragenic palindromic PCR (rep-PCR), repeat elements such as **ERIC** (Enterobacterial Repetitive Intergenic Consensus) and **BOX** elements (section 1.2.3.4). The fingerprint method is a powerful tool and allows for the study of diversity within a species at strain level, by analysing the specific fingerprints generated from distinct genomes.

The objective was to develop a semi-automated system which would allow us to study the dynamic structure of the microbial population in the soil environment. However, the method must first be developed and then optimised for the Genescan system. Two fingerprint methods, ERIC and BOX, were investigated for their potential application in the environmental studies.

8.2 Methods

8.2.1 DNA extraction method

Details of DNA extraction methods are in Chapter 2. (Extraction of bacterial DNA in 2.6.1; fast bacterial DNA extraction method, 2.6.2).

8.2.2 Standard fingerprint PCR reaction mix

50 pmol of primers

2.5 µl of x10 PCR buffer

0.5 µl of dNTPs mix [10mM conc.] (i.e. 200µmol of each nucleotide).

0.2 µl of Taq polymerase [5U/µl] (1 Unit)

1 µl of DNA solution.

18.8 µl of dH_2O (molecular biology grade)

The final volume was 25 µl.

PCR Buffer (X10)

100 mM Tris HCl pH 8.3

500 mM KCl

15 mM $MgCl_2$

(or see manufacturer's guidelines)

One drop of mineral oil was added if the Biometra I was used, and no oil if the Biometra III with heated lids was used. (For details of PCR reaction setup see 2.7.1; standard methods for gel electrophoresis are outlined in 2.6.3).

8.2.3 ERIC Primers

The DNA Sequence for the ERIC Primers are as follows:

CIRE 1: 3' **AAG TAA GTG ACT GGG GTG AGC G** 5'

ERIC 2: 3' **CAC TTA GGG GTC CTC GAA TAT A** 5'

The ERIC oligonucleotides, CIRE 1 and ERIC 2 (section 2.10.2.) were dissolved in a known quantity of dH_2O and the concentration for the PCR reaction was made up to 50 pM for each primer in 25 µl reaction mix.

8.2.3.1 ERIC PCR reaction mix

Reagents	**Volume**	**Concentration**
CIRE 1 primer	1.0 µl	(50 mols l^{-1})
ERIC 2 primer	1.0 µl	(50 mols l^{-1})
PCR buffer x10	2.5 µl	
dATPs	0.5 µl	[10 mM conc.]
dGTPs	0.5 µl	(200 µmol of each nucleotide).
dTTPs	0.5 µl	

dCTPs	0.5 µl	
Taq polymerase	0.2 µl	(1 Unit) [/ 5U/µl]
DNA solution	1.0 µl	
dH_2O	18.8 µl	(molecular biology grade)

Final volume of 25 µl.

8.2.3.2 PCR program: ERIC primers fingerprinting

Step 1	**95°C for 7 min**	
Step 2	94°C for 1 min.	Cycle from step 2 to 4 for 30 cycles.
Step 3	52°C for 1 min.	
Step 4	72 °C for 8 min.	
Step 5	**72 °C for 16 min**	Total run time: 6 hours 13 min and 12 sec.
Step 6	10 °C to cool.	

PCR products are run on a 1.5 % agarose gel. To obtain clear definition of fingerprint patterns all gels were run slowly (app. 3-4 hours) to give clear separations of bands.

Figure 8.1 shows ERIC bands of *E. coli* run on a 1.5% agarose gel for 3 hours.

8.2.4 BOX primer

The BOX Primer is described in section 2.10.4. The Box primer (supplied by Perkin Elmer) was purchased in solution and the primer concentration was unknown. The primer concentration is critical in PCR reactions and therefore the primer concentration had to be calculated before use in a reaction (see Table 8.1). The calculation was derived as follows:

8.2.4.1 Calculation of primer concentrations

The known total concentration was given at 3026.60 mg $L^{-1,}$ giving a concentration of 444.4 µM molar concentration in solution supplied. This gave a working concentration of 444 pmol μg^{-1} in the primer mix; however, the concentration of primer required in the PCR reaction had to be optimised.

Figure 8.1 ERIC bands of *E. coli* run on a 1.5% agarous gel for 3 hours

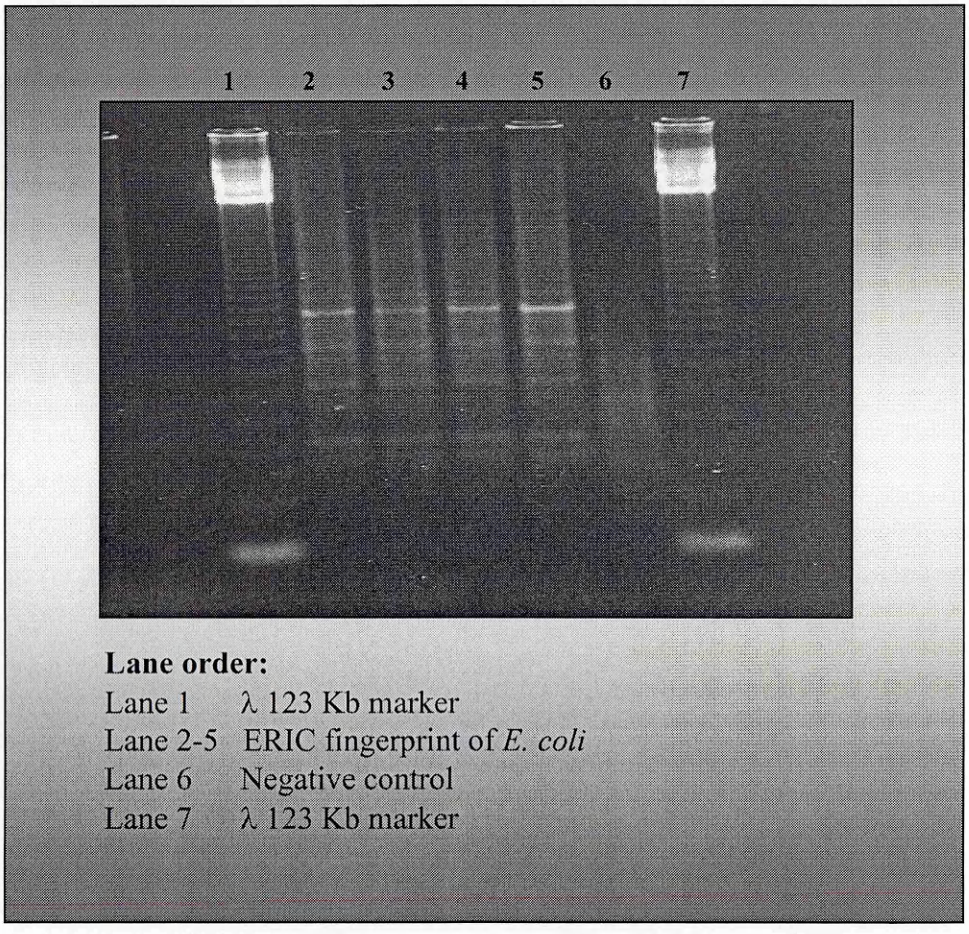


Table 8.1 Molecular weights (Mol wt) for nucleotides

	Basepairs Mol wt	Number of basepairs	BOX primer Mol Wt
A	312.2	5	1561
T	303.1	2	606.4
G	328.2	8	2625.6
C	288.2	7	2017.4
Total			**6810.4**

8.2.4.2 Annealing temperature

The annealing temperature of the PCR relates to the temperature at which the primers bind to the target DNA region. The primer melting temperature is used as an indictor of the annealing temperature for a PCR reaction. The ideal annealing temperature is the temperature at which half the primers are annealed to the target DNA region. This should be the highest temperature which gives efficient amplification of the desired product, with the lowest level of nonspecific product.

The OLIGO program: The ideal annealing temperature for a PCR reaction must be calculated to ensure a positive production of the PCR product. The temperature can be calculated by a simple computer program called the OLIGO Program (IACR-Rothamsted mainframe and various other programs accessible on the Web), or calculated manually (section 5.3.1). The annealing temperate for the BOX Primers was calculated at 65 °C. The annealing temperature had to be optimised and lowering the commuted temperature by approximately 5 °C was recommended for the first attempt. However, a range of temperatures must be used to find the annealing temperature for this application.

8.2.4.3 BOX PCR protocol

The BOX PCR fingerprint procedure, as recommended by Versalovic, was used as a template to optimise the BOX PCR (Versalovic et *al.*, 1995).

BOX PCR reaction mix

Reagents	Volume	Concentration
dNTPs	2.5 µl	1.25 mM dNTPs
Taq	0.4 µl	10 µl /ml or 2 units of *Taq*
PCR buffer	2.5 µl	
dH_2O	5.6 µl	
DNA solution	1.0 µl	
BOX Primer	1.0 µl	Concentration to be optimsed

Final volume per reaction was 25 µl.

8.3 Optimsation of fingerprint PCR

8.3 1. DNA concentrations

The Bacterial DNA was obtained as previously described (section 8.2.1) and three concentrations of DNA were tested. They were: undiluted, 1/10 dilution and 1/100 dilution of the lysed DNA solution sample. The DNA concentration was found to be optimal at 1/10 dilution of the lysed DNA solution sample. The concentration was checked on agarose gels and the DNA concentration was calculated to be between 30-50 pmol. The concentration of DNA was set at this for all further PCR experiments; due to the number of samples that were processed, the DNA concentrations were assumed to be within this range.

8.3.2 Taq polymerase enzyme

The effect of different thermostable polymerase in a PCR reaction may lead to a slightly different PCR product. This can be explained by different production criteria in the manufacturing process, which can lead to slightly different properties in the Taq enzyme, or to different properties of different enzymes. It was important to check with a standard bacterial profile to see if there was a problem in our application. Any indication of a problem can be solved by employing a standard Taq polymerase in verification assays. *E. coli* DNA was used to generate a control profile and the effect of the different polymerases on the *E. coli* ERIC profile was investigated.

8.3.2.1 Results from thermostable polymerase enzymes

Three polymerase suppliers, Boehringer Mannheim, Promega and Dynazyme™ (manufactured by Flowgen), were tested. The standard ERIC protocol was followed as set out in 8.2.3.1 and the following result was obtained: ERIC profiles varied with each of the polymerases. Figure 8.2 shows the effect that different polymerases have on the genetic fingerprint profile obtained. This emphasises the importance of a standard method and reagents, so the profiles can be compared in a large data base.

The results indicated that there was a difference in *E. coli* profiles obtained with different enzymes. Therefore, it was necessary to standardise the polymerase used in all future experiments. The Dynazyme™ and Dynazyme™ PCR buffer were chosen as they were successful in producing reproducible band patterns and also represented a cost saving. Accordingly, both Dynazyme™ polymerase and buffer were employed in all PCR reactions, unless stated otherwise.

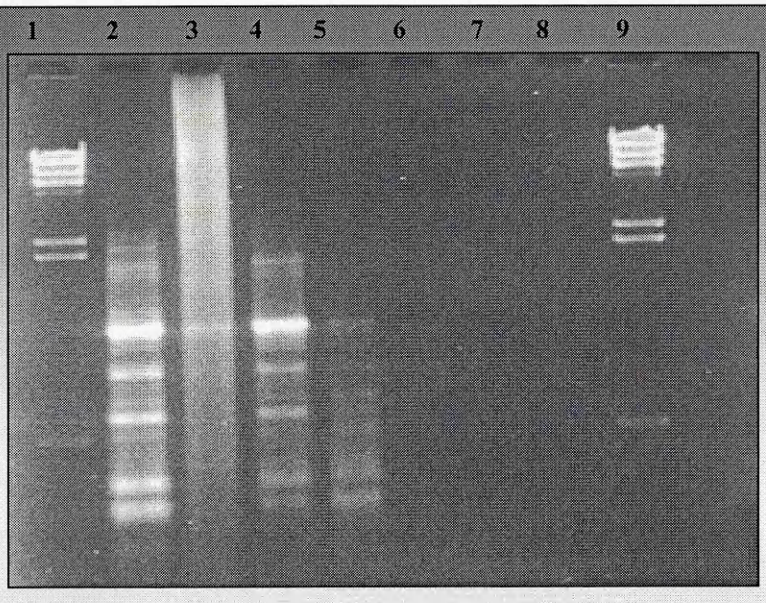
8.3.3 dNTPs concentrations

The levels of each nucleotide (dNTPs) recommended (Versalovic *et al,* 1995) were open to interpretation, so two different concentrations each dNTP- 200 μmol PCR and 50 μmol- were investigated. It was found that the correct concentration for the ERIC primers was 200 μmol for each dNTP. However, the BOX PCR reaction was shown to be problematic, indicating that other parameters had to be investigated, but the level of dNTPs for the BOX PCR was also set at 200 μmol.

8.3.4 Primer concentrations

The primer concentration for ERIC primers was recommended at 50 pmols (Versalovic *et al.,*1995). This was checked with a titration of primer concentrations and was found to be the optimum for this application.

Fig 8.2 The effect of Taq polymerases from different sources on genetic fingerprint profiles of *E. coli*



Lane order:

Lane 1	λ *Hind III*
Lane 2	Dynazyme™ Taq + Dynazyme™ PCR buffer
Lane 3	Dynazyme™ Taq + Boehringer and Mannheim PCR buffer
Lane 4	Dynazyme Taq + Promega PCR buffer
Lane 5	Boehringer and Mannheim Taq + PCR buffer
Lane 6	Promega + Promega PCR buffer
Lane 7	Blank
Lane 8	Negative control
Lane 9	λ *Hind III*

8.3.4.1 Concentration of BOX primers

The concentration of the BOX primers were investigated and optimsed. The following concentrations of primers were used; 45, 22.5, 11.2 pmols. The concentration of primers could not be optimsed, as no product was formed using the recommended procedure (section 8.2.2). Therefore, another two aspects of the PCR reaction had to be investigated: the magnesium concentrations and the PCR program.

8.3.5 Magnesium concentrations

Magnesium (Mg^{2+}) is a critical component in the PCR, as its concentration can affect the specificity and efficiency of the reaction. The Mg^{2+} concentration needs to be optimised for the particular application required. *Taq* DNA polymerase (*Taq*) is dependent upon the presence of Mg^{2+}, as it affects the fidelity (error rate) of the *Taq*. It is more error prone with excess Mg^{2+}, than with lower concentrations. In addition, the free Mg^{2+} concentration is affected by the dNTP concentration (there is equimolar binding between dNTPs and Mg^{2+}). However, the interaction between *Taq*, dNTP and Mg^{2+} should be noted, as the proofreading ability of *Taq* needs each dNTP concentration to be 200 μM to guard against nuclease activity degrading the primers.

The following Mg^{2+} Concentrations were tested for the BOX primers:

1.5mM of Mg^{2+} (standard concentration in PCR buffers with Mg^{2+})

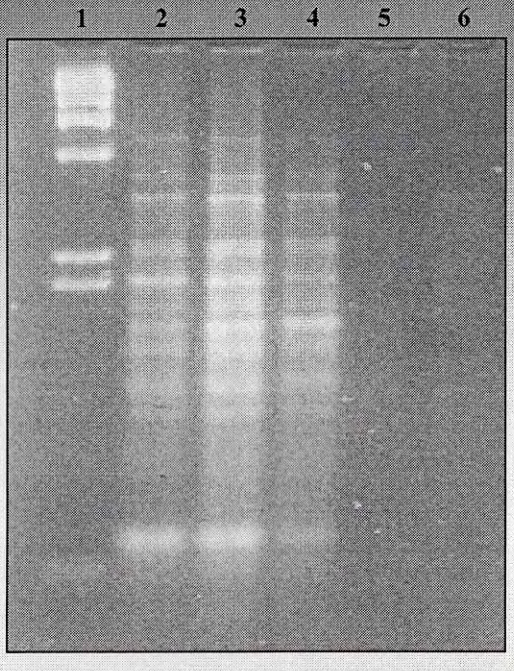
2.5mM of Mg^{2+}

3.5mM of Mg^{2+}

8.3.5.1 Result of Mg^{2+} concentrations

The result from the Mg^{2+} experiment showed that a Mg^{2+} requirement of 2.5 mM was required to obtain *E. coli* profiles with BOX primers (Figure 8.3). Therefore, the following PCR mix and PCR program was used and set as the standard required for the BOX fingerprint Method.

Figure 8.3 *E. coli* profiles with positive and negative controls, with BOX primers at varying Mg^{2+} concentration.



Lane order:

Lane 1	λ *Hind III* marker
Lane 2	*E.coli* BOX standard
Lane 3	*E.coli* BOX standard 2.5mM
Lane 4	*E.coli* BOX standard 3.5mM
Lane 5	Negative control
Lane 6	Rhizobium BOX 1.5mM

Standard PCR mix for BOX primers

Reagents	Volume	Concentration
dNTPs	2.5 µl	10 mM conc. (200 µmol of each nucleotide).
BOX Primer	1.0 µl	45 $molsl^{-1}$
dH_2O	17.5 µl	
DNA solution	1.0 µl	
Taq	0.4 µl	*Taq* polymerase [/ 5U/ul] (1 Unit)
PCR buffer	2.5 µl	
Mg^{2+}	1.0 µl	2.5 mM

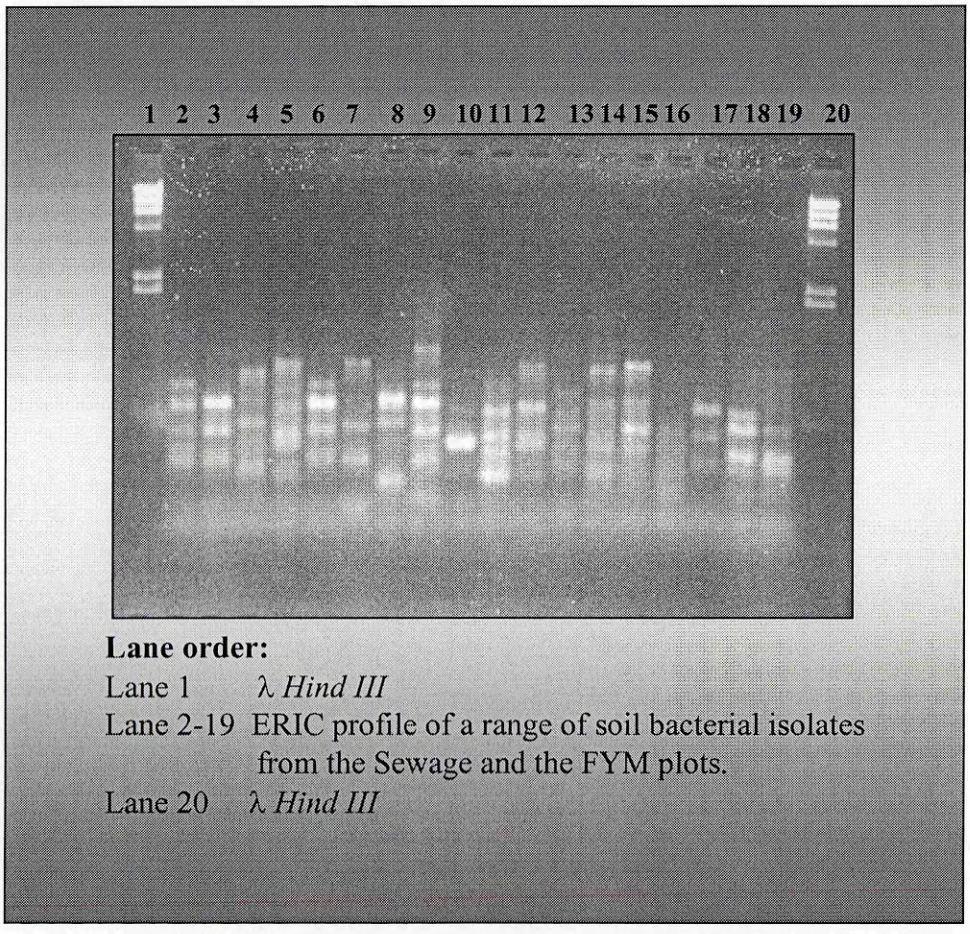
The primer annealing temperature was set at 52 °C and the BOX program was as follows:

BOX PROGRAM

Step 1	**95°C for 7 min**	
Step 2	94°C for 1 min.	Cycle from step 2 to 4 for 30 cycles.
Step 3	52°C for 1 min.	
Step 4	72 °C for 8 min.	
Step 5	**72 °C for 16 min**	
Step 6	10 °C to cool.	

PCR products are run on a 1.5 % agarose gel. To obtain clear definition of fingerprint patterns, all gels were run slowly (approximately 3-4 hours) to give clear separations of bands.

Figure 8.4 shows the ERIC profiles of a range of soil bacterial isolates from the Sewage and FYM plots.



Lane order:

Lane 1 λ *Hind III*

Lane 2-19 ERIC profile of a range of soil bacterial isolates from the Sewage and the FYM plots.

Lane 20 λ *Hind III*

8.4 Optimsation summary

DNA concentration was set at 1 /10 dilution of the DNA suspension, i.e. a concentration calculated to contain approximately 30-50 pmols of DNA.

Dynazyme enzyme and PCR buffer were appointed the standard polymerase for ERIC PCR and BOX PCR.

The concentration of each nucleotide (dNTP) was standardised at 200 μmol for both BOX and ERIC PCR.

The Mg^{2+} concentration was set at the 1.5mM of Mg^{2+} (standard concentration in PCR buffers with Mg^{2+}) for the ERIC PCR, while BOX PCR was set at 2.5mM of Mg^{2+}.

The primer concentration for ERIC primer was 50 pmols and the BOX primer concentration was optimsed at 45 mols μl.

8.5 BOX & ERIC PCR for Genescan

The fingerprint methods were optimsed and readied for application to environmental isolates, but a process objective was to develop the semi-automation of the fingerprint assimilation of environmental samples. Therefore the BOX and ERIC PCR had to be optimsed for the Genescan System (Chapter 5). The Genescan system required the bacterial PCR fingerprint to be labelled with an attached fluorescent dye which could be recognised by electronic laser in the ABI Sequencer 373 system (section 5.3.1). This allowed the fluorescently-labelled PCR fingerprint to be recognised by the Genescan system and therefore allowed the automatic sizing and analysis of the fingerprint patterns (using the developed analysis programs; Chapter 6) to be run and analysed on the Genescan system.

8.5.1 Comparing fingerprint methods

The results indicate successful fingerprints from both ERIC (Figures 8.4) and BOX methods (Figure 8.3). Figure 8.4 contains a selection of bacterial strains isolated from

the field plots, showing the range of ERIC banding patterns. This figures emphasis the difficulty with the manual sizing of ERIC profiles and demonstrated the need for the development of the Genescan system. Problems with the BOX method (outlined in 8.3.3-8.3.5) were resolved and the fingerprints subsequently obtained were consistent. However, an important difference between the two methods was the size of the PCR Products obtained. The BOX primer produced fingerprint profiles of between 1 kb to 9 kb; these would be ideal for manual sizing, but these were not required in the automated system. In contrast, ERIC primers gave PCR products in the range of 300 bp to 3 kb, a narrower range, which would cause problems with manual sizing, due to separation of band products of similar sizes.

Initial testing of labelled BOX and ERIC fingerprints indicated that with the conditions optimsed for the Genescan™ system, the best results were with products less than 2 kb. Therefore, although both fingerprint methods were operational on the system, it was decided to concentrate on ERIC primers, and optimise their working conditions. However, if at a later date the BOX fingerprint was required to confirm results, it could be adapted to the Genescan system.

8.5.2 Summary

Results from the preliminary work showed that once optimised, the BOX size fragments were larger than the ERIC primers. The ERIC profiles ran more successfully on the 373 system and the development setup of the system was therefore biased toward the ERIC primers (Chapter 5). The range of fragment sizes and the discriminatory ability of the Genescan system selected the ERIC primers.

A decision was made to optimise further only the ERIC fingerprinting in this study. Therefore I concentrated on the further optimsation of fluorescently labelling the ERIC PCR products for the Genescan system.

8.5.3 Fluorescent labelling for Genescan

The next process step was performed in conjunction with the development of the Genescan system in Chapter 5. The objective was to label successfully the ERIC-PCR

fingerprint of a known bacterial strain (*E.coli*). By using the known fingerprint profile, the effect of different concentrations of ERIC Primer labelled with 6-FAM (fluorescent blue dye) could be assessed. Another objective was to label the PCR products to a level at which the ABI Sequencer laser would be able to detect labelled products and hence automatically size them. The level of fluorescent labelling was important: too low, and the laser would be unable to register its existence; too high, and there would be interference from the high level of fluorescence.

8.5.3.1 ERIC primer labelled with 6-Fam (fluorescent blue dye)

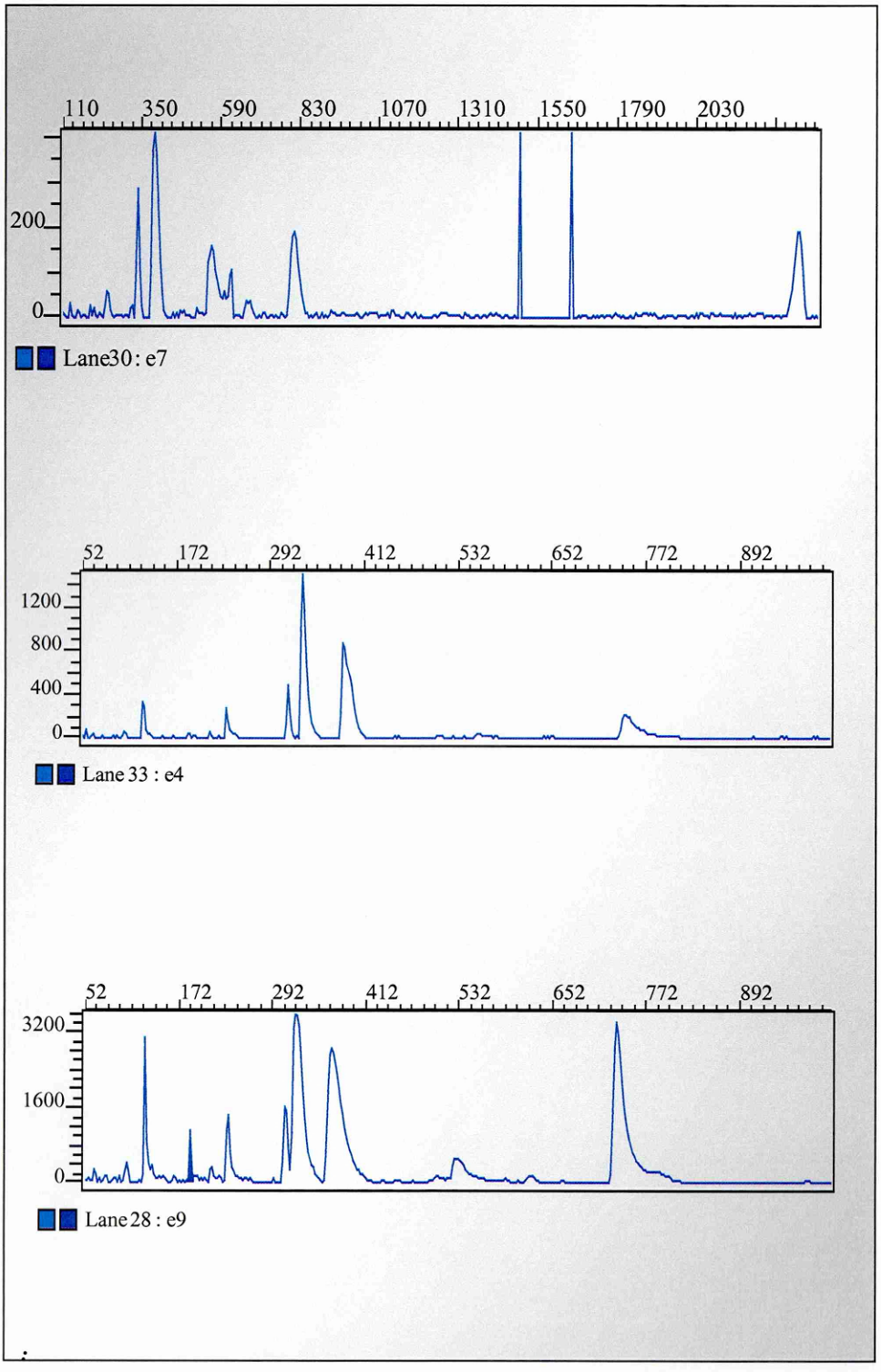
The ideal conditions for the ERIC PCR have been set out in section 8.2.3.1; the primer concentrations were set at 50 pmols in each reaction tube. The PCR reaction requires two primers ERIC 2 and CIRE 1, given that the common practice in human genetics is to use one labelled primer and one unlabelled primer (recommend by Perkin-Elmer ABI, personal communication). The following experiment was set up to verify if the conditions used in human genetics could be applied to bacterial genetics.

The profiles were run on 1.5% agarose gels, which showed that the labelled primers did not affect the PCR reaction and the band patterns obtained were the same. Therefore, the labelling had no effect on bacterial profiles. When the samples were run on the Genescan system, it revealed that the profile from one labelled primer were different from the profile with two labelled primers. It was therefore deemed necessary that for the ERIC bacterial profiles, two labelled primers were required. However, with this, the level of labelling was too high and a problem with the fluorescence levels was envisaged. Therefore the level of dual labelling had to be optimsed. Figure 8.5 shows the different combinations of labelled primer.

8.5.3.2 Titration of labelled primers

Several different protocols were employed to optimise the labelling level of the PCR products. One protocol involved starting the ERIC PCR with unlabelled primers and then adding the labelled primers after a number of cycles. It was found that by starting the reaction with unlabelled primers and adding the labelled primer at 10 cycles, an

Fig 8.5 A computer output of different combinations of labelled primer

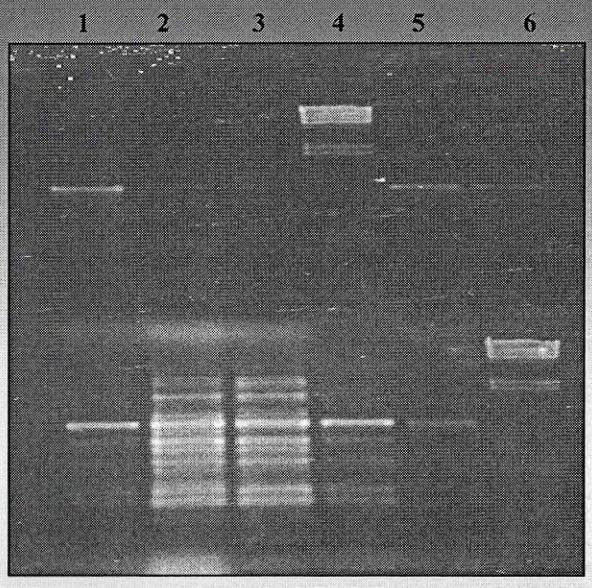


KEY ***e7: 12.5 pmol labelled***

e4: 25 pmol labelled

e9: 50 pmol labelled

Fig 8.6 The effect of different levels of labelling on the ERIC profile of *E.coli*



Lane order:

Lane 1 CIRE 1 AND ERIC 2: unlabelled + 25pmol unlabelled
Lane 2 CIRE 1 AND ERIC 2: labelled 25pmol+ 25pmol unlabelled
Lane 3 CIRE 1 AND ERIC 2: labelled 12.5pmol+ 25pmol unlabelled
Lane 4 CIRE 1 AND ERIC 2: labelled 6.5pmol+ 25pmol unlabelled
Lane 5 CIRE 1 AND ERIC 2: labelled 3.5pmol+ 25pmol unlabelled
Lane 6 λ *Hind III*

acceptable level of labeling was obtained. This protocol was only feasible with a small number of samples as the PCR reaction was too complicated for larger scale sampling. The increased number of steps increased the likelihood of cross-contamination of the DNA samples. Therefore the PCR was standardised with the least number of steps possible. The protocol set out in Figure 8.6 shows the uses of various levels of unlabelled primers, versus labelled primer, but all primers were added at the start of the reaction with the standard conditions set out in 8.2.3.1.

The products were run on a 1.5% agarose gel and, as can be seen in Figure 8.6, the two primer concentrations that produced sensitive profiles were: (b) CIRE 1 & ERIC 2: labelled 25 pmol + 25 unlabelled; and (c) CIRE 1 & ERIC 2: labelled 12.5pmol+25 unlabelled.

However, when the samples were run on the Genescan system, the acceptable level was found to be (b) unlabelled 25 pmol, and 25 pmol labelled primers. All further labelled PCR reactions contained this level.

8.6 Discussion

The main process objective discussed in this chapter was to standardise all fingerprinting procedures and to optimise the genetic profiling of bacterial fingerprints. The PCR products needed to be analysed on an ABI Sequencer using Genescan software to produce a database of ERIC fragment sizes, allowing isolates to be compared and assigned to groups.

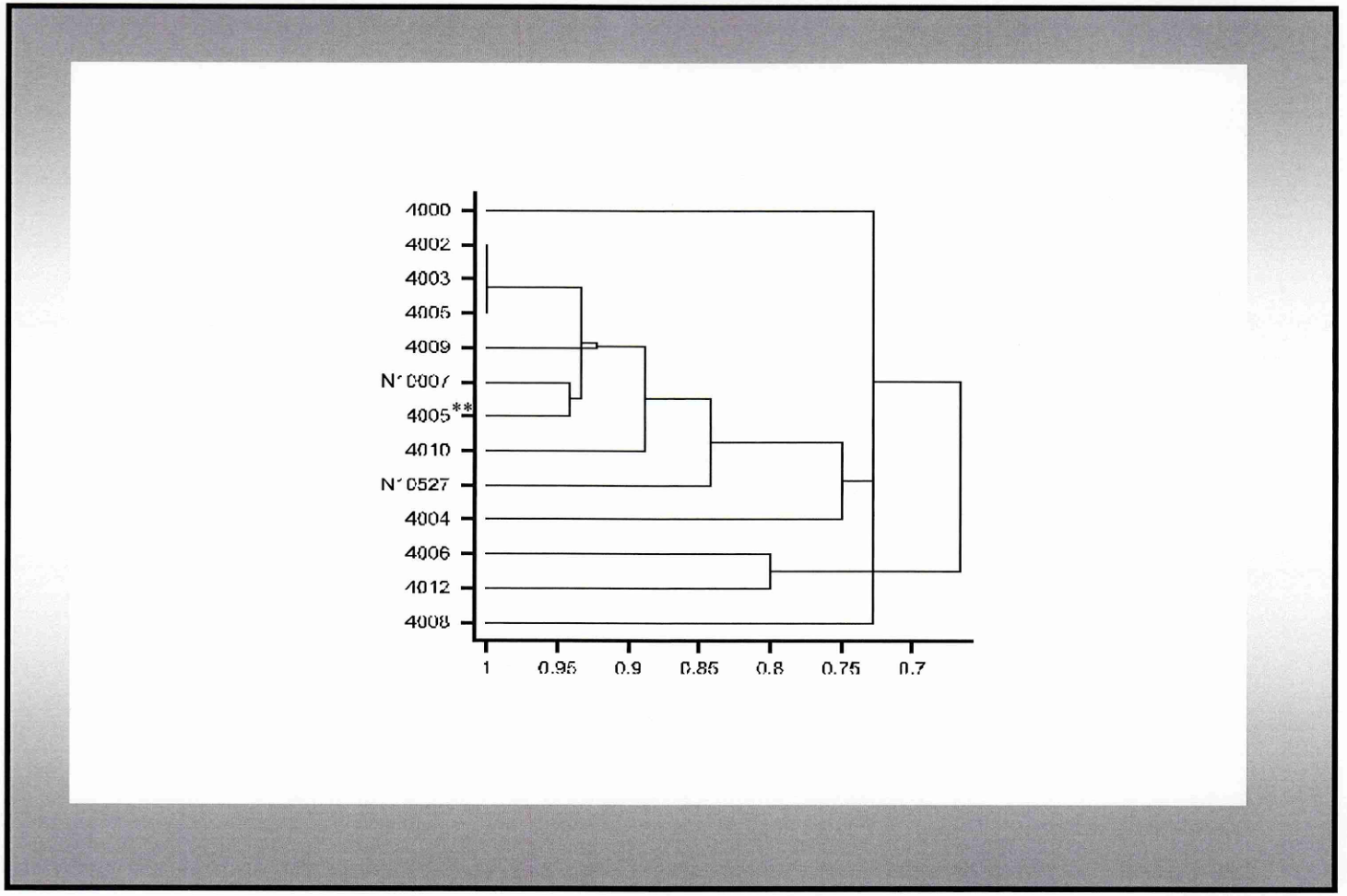
Two fingerprint methods were investigated and fully standardised to a position that they could be developed for further use on the Genescan system.

The ERIC PCR was successfully optimised for Genescan with all procedures set up. The final stage of the project was to investigate the genetic diversity of bacteria by using their unique PCR profiles.

8.7 Conclusions

The ERIC PCR fingerprint method was standardised and optimised for use on the Genescan system using the 373 ABI Sequencer. This allowed the investigation into the genetic diversity of the culturable pseudomonads from the two soils to be undertaken, which is the subject of the next chapter.

Figure 9.1. Phylogenetic tree (UPGMA) of reference strains



Chapter Nine:

Genetic Diversity in FYM and Sewage

9.1 Introduction

This chapter investigates genetic diversity by studying the distribution of genetic fingerprint profiles within the soil microbial populations of the two plots. It allows us to profile the key components of a specific bacterial sub-population in greater depth than is possible by using other methods which are essentially a simplistic or 'one-dimensional' measures (i.e. colony counts, or catabolic activity). Indeed, the previous chapters have aptly demonstrated the limitations of these techniques.

9.1.1 Genetic diversity

Bacterial diversity is not straightforward, as prokaryotes are difficult to define in conventional classification (Rosselló-Mora & Amann, 2001). In bacterial genetic diversity, a decision has to be made at what point the DNA of individual clones, descended from a common parental cell, has evolved to be sufficiently different that they can be recognised as separate types, biovars, species or genera. The acquisition of DNA elements from other bacteria, rearrangement of genomic DNA, and the accumulation of point mutations will all contribute to such changes over time. Taxonomic and chemotaxonomic studies provide the basis of conventional bacterial classification, with DNA-DNA hybridization studies providing quantitative information based on genetic relatedness (within a species, strains share 70% or greater DNA sequence similarity [Wayne *et al.,* 1987]). The phylogenetic approach to classification (section 1.2.1.3) has been widely used and the 16S rRNA gene sequence is frequently exploited. Bacteria have several mechanisms for transferring genetic material horizontally between distantly-related, even unrelated species, and there are cases where 16S rRNA genes may have been exchanged (Young 2001, Fox *et al.,*1992). Although bacterial species are difficult to define, the basic clonal unit of individual cells, descended relatively recently from one parental cell, can be identified using DNA fingerprinting methods. Similarly, fingerprints of phenotypically-similar organisms will provide some information on genetic relatedness.

9.1.2 *Pseudomonas* genus

Bergey's manual of Systematic Bacteriology characterised the Family Pseudomonadaceae as polarly flagellated Gram-negative rod-shaped aerobic bacteria (Bergey, 1993). The genus *Pseudomonas* is very heterogeneous and contains a large number of species; indeed, this broad and essentially vague phenotypic definition has allowed it to become a dumping ground for *Pseudomonas*-like species. These were classified as one genus, *Pseudomonas*.

It is now recognised that they fall into several different taxonomic groups, spanning different subclasses of the Proteobacteria (Anzai *et al.,* 2000). A revised description of the genus *Pseudomonas* has now been proposed by Anzai, according to the phylogenetic studies based on 16S rRNA sequences, chemotaxonomic studies, and further taxonomic studies. For the purpose of this study, therefore, the pseudomonad group is examined under the broadest definition of *Pseudomonas* genera.

9.1.3 Genetic fingerprinting of *Pseudomonas*

By studying the complex relationship of discrete soil sub-populations with genetic fingerprinting techniques, we can leverage improved understanding of population dynamics, particularly within the pseudomonad group. Population diversity can be studied at different levels, with various degrees of relevance to the genetic relationships between the isolates studied.

Existing research describes multiple methods for studying genetic diversity, but three fingerprint methods were selected for this thesis. These were:

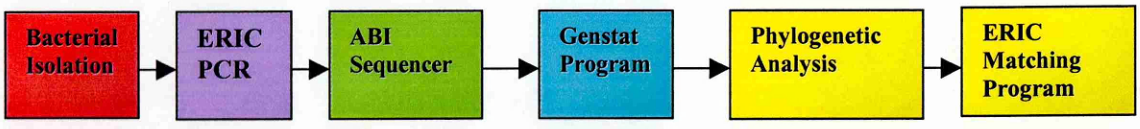
16S rRNA Universal Primer Profiles (9.3.2.1)
16S rRNA Pseudomonas-Specific Primer Profiles (9.3.2.2)
ERIC PCR Fingerprinting Profiles (9.3.3)

These methods were reviewed in Chapter 1 (sections 1.2.3.2; 1.2.3.3) and validated experimentally on environmental reference strains (Table 9.1) to validate the potential of each of these fingerprint techniques.

16S methods must first be applied to determine the degree of similarity between isolates; ERIC can then be used to group related individuals. Whilst this study has gathered profile data derived from all three techniques, the main concern of this chapter is concentrated on the evaluation of ERIC PCR profiles from **all individuals** isolated from both soil plots from the Woburn Garden Experiment (section 2.1), at sampling times described above (Table 3.1).

9.1.4 Role of ERIC profiling

The schematic below illustrates where ERIC profiles and phylogenetic analyses fit, as indicators of genetic shifts within bacterial communities, into the overall methodology used in this thesis. These methods are sequential because each analysis is dependent on outputs from the previous step.



i) Bacterial isolation: The population chosen for analysis, selected on PSA media, was an agriculturally important group implicated in healthy plant growth-namely the fluorescent pseudomonads. (However, account was taken that bacteria *other* than those related to the genus '*Pseudomonas*' may grow on PSA media.)

ii) DNA extraction and ERIC PCR fingerprinting: Using ERIC-PCR (de Bruijn *et al.*,1995), all isolated colonies were fingerprinted and subsequently analysed (see Chapter 8).

iii) ABI sequencer and genescan: The bacterial ERIC-PCR fingerprints were run on the ABI Sequencer, allowing subsequent analysis using Genescan (Chapters 4 and 5).

iv) Genstat programs: Genescan™ software was used to process outputs from the ABI sequencer and then run on a Microsoft Access database program, to convert ERIC data into a format permitting comparison of bacterial isolates and subsequent assignment into groups (Chapter 6).

v) Phylogenetic analysis: The PHYLIP version 3.57 software package UPGMA (*Unweighted Pair Group Method using Average linkage,* Felsenstein, 1996) was used to display the relationship between the bacterial isolates (Chapter 4).

vi) ERIC matching program: This was a simple matching program on distribution of band products, which allowed the size of band segments to be set in a Microsoft Access database program (section 6.4).

The 16S ARDRA was used, in addition to the ERIC fingerprint method, to examine genetic diversity. ARDRA (amplified ribosomal DNA-restriction analysis, [Vaneechoutte *et al.*, 1993]) is a simple classification of bacteria, based on the restriction digest profiles of the 16S rRNA PCR product. The 16S profiles were manually analysed and the details highlighted in Chapter 2.

9.2 16S rRNA methods

Table 9.1 contains a list a of environmental isolates from the Rothamsted rhizosphere culture collection, isolated initially from wheat roots on PSA, and identified using the API-NE system, together with two soil pseudomonads from the National Collections of Industrial, Food and Marine Bacteria (NCIMB), Scotland. They are used as reference strains in this study.

Table 9.1 Collection of reference strains used to verify fingerprinting methods

Rothamsted rhizosphere collection	
Strain Numbers*	Putative Strain Names based on API
RSM 4000	*Pseudomonas (P) fluorescens*
RSM 4002	*P. chlororaphis* or *B. cepacia*
RSM 4003	*P. chlororaphis*
RSM 4004	*P. fluorescens*
RSM 4005	*Agrobacterium (A). radiobacter*
RSM 4005 (LB)**	*Agrobacterium (A). radiobacter*
RSM 4006	*P. chlororaphis* or *B. cepacia*
RSM 4008	*P. fluorescens* or *chlororaphis*
RSM 4009	*P. chlororaphis*
RSM 4010	*P. fluorescens*
RSM 4012	A. *radiobacter*
NCIMB reference stains	
NCIMB 10527	*P. fluorescens*
NCIMB 1007	*P. putdia*

**all strains were grown up in low nutrient 1/10 TSA broth*

*** due to poor growth, high nutrient LB was used*

9.2.1 Sample preparation

Cultures were grown in 1/10 TSA (RSM 4005 grown in LB as well) and were then lysed to release DNA and amplified with specific primers. The PCR products for the ERIC fingerprints were analysed using the developed GeneScan System (Chapter 5-6). 16S Universal (Woese, 1987) and 16S Specific (Widner *et al.,* 1998) techniques are described in detail in Chapter 2. The PCR amplified 16S rDNA products were digested with a selection of restriction enzymes (RE) and the fragments were run on 3 % Nusieve agarose gels. The restricted products generated a distinct restriction pattern. These species-specific patterns (mostly) reflect the conserved character of the rRNA genes (Appendix C).

9.3 Results from reference strains

9.3.1 Success of PCR reactions

The outcome of each PCR reaction is recorded in Table 9.2. All reference strains were successfully amplified by the ERIC primers. The 16S rRNA Primers (both Universal and Specific) failed to yield a product for the following isolates: RSM 4005, RSM 4006, RSM 4012, RSM 4005*. The experiment was repeated and result confirmed. The strain RSM 4009 was amplified by the 16S rRNA Universal primers, but not the Specific Primers, indicating that it may not be a pseudomonad strain. The Specific Primers, however, successfully amplified the remaining isolates.

Table 9.2 Summary of amplified PCR products using three identification primers

Strains	PRIMERS		
	16S Universal	16S Specific	ERIC
RSM 4000	+	+	+
RSM 4002	+	+	+
RSM 4003	+	+	+
RSM 4004	+	+	+
RSM 4005	-	-	+
RSM 4006	-	-	+
RSM 4008	+	+	+

RSM 4009	+	-	+
RSM 4010	+	+	+
RSM 4012	-	-	+
NCIMB 10527	+	+	+
NCIMB 1007	+	+	+
RSM 4005 (LB)	-	-	+

- = no PCR product; + = PCR product

9.3.2. 16S rRNA ARDRA

9.3.2.1 16S rRNA Universal

The Universal profile of the *P. fluorescens* strain NCIMB 100527 and the strains RSM 4000, RSM 4002, RSM 4003, RSM 4004, and RSM 4010 were very similar, except for with *Alu I.* (Table 9.3). The Universal profile for strain RSM 4009 was dissimilar to the other isolates. *P. putida* had no similarity to the Rothamsted rhizosphere strains.

Table 9.3 ARDRA profile of reference strains, groups A to R are arbitrary groups designed to group identical profiles.

Strains	Universal 16S			Specific 16S		
	Rsa I	*Hinf I*	*Alu I*	*Rsa I*	*Hinf I*	*Alu I*
RSM 4000	I	P	M	A	C	F
RSM 4002	I	P	M	A	C	F
RSM 4003	I	P	M	A	C	F
RSM 4004	I	P	M	A	D	F
RSM 4008	I	P	O*	A	D	F
RSM 4009	K	Q	M	N/A	N/A	N/A
RSM 4010	I	P	N	A	C	F
P. putida NCIMB 1007	J	R	L /M*	B	E	G /F*
P. fluorescens NCIMB 10527	I	P	L /M*	A	C	G /F*

** classification unsure; see 9.3.2.3*

9.3.2.2 16S rRNA Specific

P. putida (NCIMB 1007) is a pseudomonad-type strain (ie, it is an authenticated culture collection *P. putida*) but had no similarity to the rest of the strains. The isolate RSM 4009 was not amplified with the specific primers; hence, by our preliminary definition, it is not considered eligible for classification as a pseudomonad strain. Using the *P. fluorescens* strain NCIMB 10527 as a reference profile, the following strains- RSM 4000, RSM 4002, RSM 4003, RSM 4010- were shown to be similar to *P. fluorescens*-type strains.

The Specific profiles identify three groups:

Group 1: *P. fluorescens* type strains, containing: NCIMB 10527, RSM 4000, RSM 4002, RSM 4003 and RSM 4010

Group 2: *P. fluorescens*-like strains, RSM 4004 and RSM 4008

Group 3 : *P. putida,* with no relationship to *P. fluorescens*-like strains.

This shows that strains such as RSM 4002, identified as *Burkholderia* by API, are actually *Pseudomonas*, and closely related to *P. fluorescens.*

9.3.2.3 Alu I

The *Alu I* restriction digest profiles for specific primers shows NCIMB 10527 (group G/ F*), and strains RSM 4000, RSM 4002, RSM 4003,RSM 4010 (Group F), are similar. With the Universal primers NCIMB 10527 (group L/M*) and the strains RSM 4000, RSM 4002, RSM 4003, RSM 4004 (Group M) are again similar, but not identical. The differences may possibly be explained by new sequence data (I. Clark, personal communication, 2001) showing that the 16S of strain RSM 4002 has a one base pair substitution, compared to the type strain of *P. fluorescens.* The sequences were analysed using PHYLIP and the base change was shown to affect one of the 7 *Alu I* restriction sites (data not shown).

9.3.2.4 Summary

P. putida and isolate RSM 4009 had no similarity to the rhizosphere pseudomonad reference strains. The similar profiles show that *P. fluorescens* (NCIMB 10527) and strains RSM 4000, RSM 4002, RSM 4003, RSM 4010 can be grouped together. While the strains RSM 4004 and RSM 4008 are closely related to this group, the different *Hinf*

I profiles of 16S PCR products mean that they are sufficiently different to merit classification in a separate sub-group, related to *P. fluorescens*. The results also indicate that some of the initial identification of the field isolates, using API, was misleading. However, the fluid phylogenetic relationships within pseudomonads have always caused problems in identifying and classifying *Pseudomonas* and related genera (Widmer *et al.*, 1998).

9.3.3 ERIC PCR fingerprinting profiles

Profiles were obtained for all the reference strains listed in Table 9.1 and were subsequently analysed using the modified GeneScan system (Chapters 5-7). The fingerprint data is displayed in two formats.

9.3.3.1 Phylogenetic tree results

The phylogenetic tree (Figure 9.1) is a UPMGA tree which shows the genetic relationship between the reference stains. At the 70% similarity level, there is one large group containing RSM 4001, RSM 4002, RSM 4003, RSM 4005, RSM 4010, NCIMB 10527, RSM 4004, NCIMB 1007, RSM 4005*. This is the *P. fluorescens* group and it agrees with the 16S profiling results. The *P. putida* was placed into a *P. fluorescens* sub-group, but more data, or more isolates, are needed to confirm its position in the tree.

The ERIC profiles provides more information, from the 70% level to the 100% level, showing the complex relationship between strains. However, to display more simply inter-species relationships, the ERIC matching program was also employed.

9.3.3.2 ERIC matching program results

Using ERIC fingerprinting, and its associated programs, a profile was obtained for each of the reference strains. These profiles were then grouped using the ERIC Matching Program (J. Antoniew, personal communication, 2000; Appendix A). The ERIC profile data, displayed in Table 9.4 (a and b), illustrates that the strains are all related and the potential for ERIC to provide better insights into intra-species variation.

Table 9.4a Number of bands in reference strain ERIC profiles

Strains	Number of ERIC Bands
RSM 4000	8
RSM 4002	9
RSM 4003	8
RSM 4004	4
RSM 4008	7
RSM 4009	8
RSM 4010	9
***P. putida* NCIMB 1007**	7
***P. fluorescens* NCIMB 10527**	8

Table 9.4b. Relationship between strains' ERIC profiles using ERIC matching program

ERIC profiles of reference strains		
Identical	**One Mismatch**	**Two Mismatches**
RSM 4005	RSM 4005	RSM 4009
RSM 4002	RSM 4002	RSM 4005
RSM 4003	RSM 4003	RSM 4002
		RSM 4003
		NCIMB 10527
		RSM 4010
NCIMB 10527	NCIMB 10527	RSM 4005
RSM 4010	RSM 4010	RSM 4009
		RSM 4002
		RSM 4003
		NCIMB 10527
		RSM 4010
		NCIMB 10007
		RSM 4004
RSM 4006	RSM 4006	RSM 4006
		RSM 4012
		RSM 4008
RSM 4012	RSM 4012	
RSM 4008	RSM 4008	
RSM 4009	RSM 4009	

The matching program was set at three different levels: identical profiles, profiles with one band mismatch, and profiles with two band mismatches. The following results were obtained:

i) Identical profiles: NCIMB 10527 and RSM 4010 were found to be identical and grouped together. Another group contained RSM 4005, RSM 4002 and RSM 4003, which were also identical.

ii) One band mismatch: NCIMB 10527 and RSM 4010 were grouped together, as were RSM 4005, RSM 4002 and RSM 4003. Strains RSM 4006, RSM 4012, RSM 4008, and RSM 4009 were still ungrouped.

iii) Two band mismatches: With two band mismatches all the isolates are formed into three groups. However, as the strains are all thought to be related, it is possible for a strain to be assigned to two groups.

9.3.3.3 ERIC summary

These results strongly indicate that the ERIC profiles provide more precise information on the relationships of the reference strains and showed relationships where none were apparent from the ARDRA study. The ERIC PCR Fingerprinting amplifies differences between closely related strains and enables us to form them into groups.

9.3.4 Genetic analysis of reference strains

Using these methods, the reference strains appear to be closely related. The rhizosphere isolates, selected on PSA, all appear to belong to the genus *Pseudomonas*, despite the API classification. The relationships between strains, and the question of the base substitution in the 16S rRNA sequence, meant that ARDRA did not supply adequate information on the population structure of the reference strains. The ERIC profiles provided more information on the reference strains, inferred a population structure, and showed a relationship between them. One possible explanation may be that 16S amplifies only a limited part of the gene, while ERIC amplifies sections of the *entire* genome (plus any plasmid with ERIC primers binding sites). This seems to provide a more representative understanding of the relationships between isolates, leading to a better structural overview.

The Rothamsted culture collection isolates were selected on PSA and classified using the API identification system (G. Ross *personal communication* 1997). The recent re-classification of pseudomonads based on 16S rRNA sequences (Anzai *et al.,* 2000) shows that the older classification based on metabolic phenotypes may be misleading and explains the misclassification of Rothamsted field isolates using API, however they still remain a closely-related group. It may also indicate that the API system is not suitable for soil isolates as it was designed originally for identification of medical bacterial isolates. Therefore, it is still relevant to treat the environmental isolates, which were also isolated on PSA, as a related group.

To verify these preliminary results, a number of unknown soil isolates was studied to assess possible population shifts due to contamination with heavy metals.

9.4 Genetic analysis of bacterial field isolates

The 16S rRNA and ERIC PCR fingerprinting techniques were used with the environmental isolates to verify the conclusions of the preliminary study, which indicated that the ERIC profiles provided more strain information. The putative pseudomonad isolates, harvested in October 1998, were selected to be investigated by 16S ARDRA and ERIC PCR Fingerprinting. (The methods for DNA extraction are described in 2.2 and DNA fingerprinting methods are described in 2.10.)

9.4.1. 16S rRNA results

16S Specific Primers

The discriminatory ability of the Specific Primers (2.8.2.1) was investigated in October 1998 for pseudomonad isolates from the Woburn Market Garden Experiment. Only 61% of the isolates yielded a ribosomal profile with the 16S specific primers. The ARDRA profiles, shown in Table 9.5, illustrate a varied group of pseudomonad isolates which have shown very little similarity.

Table 9.5. 16S profiles of the bacterial isolates harvested in October 1998 with arbitrary classification of profiles from A to S, which allows for identical profiles to be group.

	Isolate name	*Cfo I*	*Hae III*	*Rsa I*	Group
1	fn 1	A	A	A	1
2	fn 2	A	A	A	1
3	fn 3	A	A	B	2
4	fn 4	B	A	A	3
5	fn 6	B	A	A	3
6	fn 8	N/A	A	C	4
7	fn 11	C	A	E	5
8	fn 12	D	I	F	6
9	fn 13	E	I	G	7
10	fn 14	D	I	G	7
11	fn 13	C	I	H	8
12	fn 16	B	I	I	9
13	fn 17	F	A	E	10
14	fn 19	G	A	J	11
15	fn 20	H	A	A	12
16	fn 21	I	A	B	13
17	fn 22	J	A	K	14
18	fn 23	K	I	L	15
19	fn 24	L	I	L	16
20	fn 27	J	A	L	17
21	fn 28	M	IB	L	18
22	fn 29	N/A	I	G	7
23	fn 30	D	I	G	7
24	sn 1	N/A	B	N/A	20
25	sn 2	N	B	C	21
26	sn 7	B	B	C	4
27	sn 8	N	A	C	4
28	sn 9	N	B	C	4
29	sn 11	P	IA	G	22
30	sn 12	J	IB	M	23
31	sn 13	N/A	N/A	N/A	24
32	sn 14	Q	IA	N	25
33	sn 15	N/A	IG	N/A	26
34	sn 18	B	I	O	27
35	sn 17	G	A	P	28
36	sn 18	L	I	P	29
37	sn 19	L	I	M	29
38	sn 20	L	I	M	29
39	sn 21	G	IC	Q	30
40	sn 22	R	IC	R	31
41	sn 23	A	IC	A	32
42	sn 24	A	IE	A	32
43	sn 25	A	I	S	33
44	sn 29	A	IC	A	34
45	sn 30	A	IF	A	34
46	sn 31	A	E	A	34

Sn = isolate from the Sewage Plot; Fn = isolate from the FYM plot

Most of the bacterial field isolates were ungrouped, but six specific groups were nonetheless identified. The groups consisted of the following:

- **Group 1** with two isolates (FYM)
- **Group 3** (FYM) with two isolates
- **Group 4** (mixed) with four
- **Group 7** with two (FYM)
- **Group 32** with two (Sewage)
- **Group 34** with three isolates (Sewage)

Because results were complicated, it was hard to draw firm conclusions about population structure.

As previously reported, these restriction digest profiles, particularly those matching profiles obtained from the *Hae III* digest, were successful in discriminating between field isolates (Widner *et al.*,1998). Hence, the *Hae III* profiles were condensed and reanalysed (Table 9.6). This indicates that there was more grouping within FYM isolates compared to Sewage isolates.

Table 9.6 *Hae III* profiles of the October 1998 isolates with arbitrary classification of profiles from A to IG, showing the identical profiles grouped together.

Profile	**A**	**B**	**E**	**I**	**IA**	**IB**	**IC**	**ID**	**IE**	**IF**	**IG**
FYM	13	0	0	10	0	1	0	0	0	0	0
Sewage	2	5	1	5	3	1	4	1	1	1	1
Total	*15*	*5*	*1*	*15*	*3*	*2*	*4*	*1*	*1*	*1*	*1*
%	**31**	**10**	**2**	**31**	**6**	**4**	**8**	**2**	**2**	**2**	**2**

All FYM isolates fell into just three groups, all of which contained sewage isolates. The remaining sewage isolates fell into three main groups, comprising six unique profiles (ungrouped). This trend indicates a more closely related population within the FYM population structure, with the Sewage populations showing signs of being more diverse.

9.4.2 ERIC PCR fingerprinting

Environmental isolates were analysed using standard parameters (section 6.4.1); with these broader parameters, no mismatches were permitted, so only matching ERIC profiles were considered. The ERIC Matching Program displayed satisfactory relationships between the 76 successful isolates (39 Sewage; 37 FYM isolates) obtained from the October 1998 sampling. Of these, 56% of Sewage isolates and 32% FYM isolates remained ungrouped. Only 34 % of the total isolates fell into 15 groups: 7 mixed; 6 groups containing FYM isolates only; and one Sewage isolate only (Table 9.6). The populations in the two plots had similarities shown by the mixed groups, but there was also a high percentage of ungrouped isolates, revealing a more complex population structure. There was more structure in the FYM plot (with 68% forming groups) compared with the less structured sewage plot (depicted by the 56 % isolates remaining ungrouped).

The results may reflect the complex nature of the soil environment. Alternatively, they could possibly be a function of skewed sampling (i.e. biases representation of the bacterial populations). One sampling point is insufficient to draw firm conclusions on population structure from such a diverse population. Clearly, temporal samplings are needed and the use of multiple samplings should permit genetic relationships, and in turn the genetic structure of these populations, to be inferred. That would allow for a more effective assessment of population shifts due to heavy metal contamination.

9.4.3 Summary

The 16S ADRDA had only a 61% success rate with the isolates, but it did provide an indication that the FYM population had more structure, indicating more clonal relationships compared with the more diverse Sewage population. ERIC profiling proved a successful technique with soil isolates and showed similarities between the two populations. However, it also indicated more structure in the FYM population alongside more diversity in the Sewage population.

The study of the three different fingerprint techniques shows that each is a valid way of bacterial fingerprinting. However, with closely related strains the 16S technique

(Universal and Specific) were not as discriminatory between strains as the ERIC technique. The 16S profiling in theory discriminates between species (within a genus), but the discriminatory value was low in this study. To increase the discriminatory value would necessitate more restriction enzyme profiling, but this system is manual and highly labour intensive, making it less feasible in any large scale environmental study. The ability of ERIC to distinguish between species within a defined genus (i.e. *Pseudomonas*, or closely related isolates grown on PSA), has been shown to be of significant value in this study.

9.4.4 Discussion and conclusion

The *Pseudomonas* genus classification was once based on phenotypes alone; now the clustering of pseudomonad isolates is based mainly on 16S rRNA studies (Anzai *et al.*, 2000). Research on 16 ribosomal fingerprinting techniques initially indicated they would be able to classify phylogenetically diverse pseudomonad species (Van Elsas *et al.*, 1997). However, the results of this preliminary study have not corroborated their ability to distinguish between related species, let alone different strains. The conclusion drawn from the study of the environmental reference strains was that the pseudomonad group might be compiled into one, or a few related groups, using these techniques.

A structured population can be defined loosely as one where strains are related to each other, either through replication (i.e. clones are generated and a few base changes occur), but overall genetic organization remains similar; or by gene transfer. The 16S rRNA gene sequence has been used to assess population structure, but there have been reports that the 16S rRNA sequence may be insufficient to guarantee species identity (Fox *et al.*,1992). The 16S rRNA results showed that the Rothamsted culture collection reference strains had some sequence variation with the 16S rRNA region (I. Clark, personal communication, 2001). This raises two points of interest on the uses of the 16S rRNA sequences:

(1) the question raised by Fox on the species identity ability of the 16S (Fox *et al.*, 1992)

(2) if 16S only amplifies a limited part of the gene, does it provide sufficient information to give a more representative idea of relationships between isolates, leading to a better structural overview. For example, strains could exchange large regions of genetic information and retain 16S rRNA genes. Conversely, the region with the rRNA genes could be transferred.

The vast biodiversity of the soil environment can only be practically handled as an ordered structure, artificial or not. Fox (1992) has challenged the established convention that an isolate is forced either to belong or not belong to a given group, proposing instead that there are particular *degrees* of group membership. Significantly, this implies that bacterial strains do not have to belong to a single cluster. Nevertheless, the concept of *partial group membership* (i.e. membership of more than one cluster) needs to be considered alongside the application of *phylogenetic logic* to bacterial population structures, which in contrast permits an isolate to be a member of one group but makes no allowances for any deviation from such a model. In essence, therefore, the evidence demonstrates that the ERIC PCR Fingerprint technique may be the more versatile and practical approach to study genetic diversity.

9.5 Comparison of population structure using ERIC PCR

9.5.1 ERIC PCR fingerprinting

Three temporal samplings were taken in March '97, March '98 and October '98 from both FYM and Sewage Plots. The putative pseudomonad isolates were isolated in a standard way, and genetic diversity assessed using ERIC PCR fingerprinting.

The proprieties of the soils sampled were tested to ensure that a possible heavy metal or organic matter 'hot spot' was avoided. The soil pH was 6.7; both carbon and nitrogen percentages and levels of heavy metals are set out in Table 9.7. This table shows that the heavy metals in the FYM plot are at background levels, while the Sewage plot has elevated levels, which were nonetheless around the limits set by EU guidelines. So, it is

important to stress that the acute effects of heavy metals have not been studied, but rather the more subtle effects or long-term exposure to elevated levels of heavy metals.

Table 9.7. Soil properties of FYM and sewage plots (Carbon, nitrogen, heavy metal)

Element:	Sewage	FYM	Element	Sewage	FYM	UK limits
%N	0.185	0.163	Zinc	234.1	94.5	300
			Copper	69.7	22.7	135
%C	2.141	1.821	Nickel	23.9	14.6	75
			Cadmium	6.35	1.44	3
			Chromium	99.1	43.5	400
			Lead	75.9	34.5	300

9.5.2 Population structure

Using the ERIC technique, populations of the FYM and Sewage plots are displayed in Figures 9.2a and 9.2b, with October 1998 shown in Table 9.5. The total number of isolates screened was over 1000 with 968 isolates purified and prepared for the PCR reaction (PCR success rate varied, but averaged 50%). The number of isolates successfully analysed using standard analysis parameters had a success rate of 57.2%. The expectance criteria for profiles in this thesis was set extremely high for development purposes and all parameters were very stringent; these can be modified for future studies. The result from the 1997 sampling shows clear differences in population structure between plots. The increase in isolate numbers in March 1998 makes the dendrogram difficult to interpret, and with increasing numbers of samples, it was not possible to produce dendrograms, so other presentational methods were used.

9.5.2.1 Results

Dendrograms drawn using UPGMA on profiles from March 1997 appear to indicate a more structured population in the FYM plot than in the Sewage plot. Hence the Sewage population appears to have greater population diversity (i.e. each isolate is less related to the others) compared to the FYM plot. However, when this was repeated the

Figure 9.2a Population structure between plots (samples from 1997)

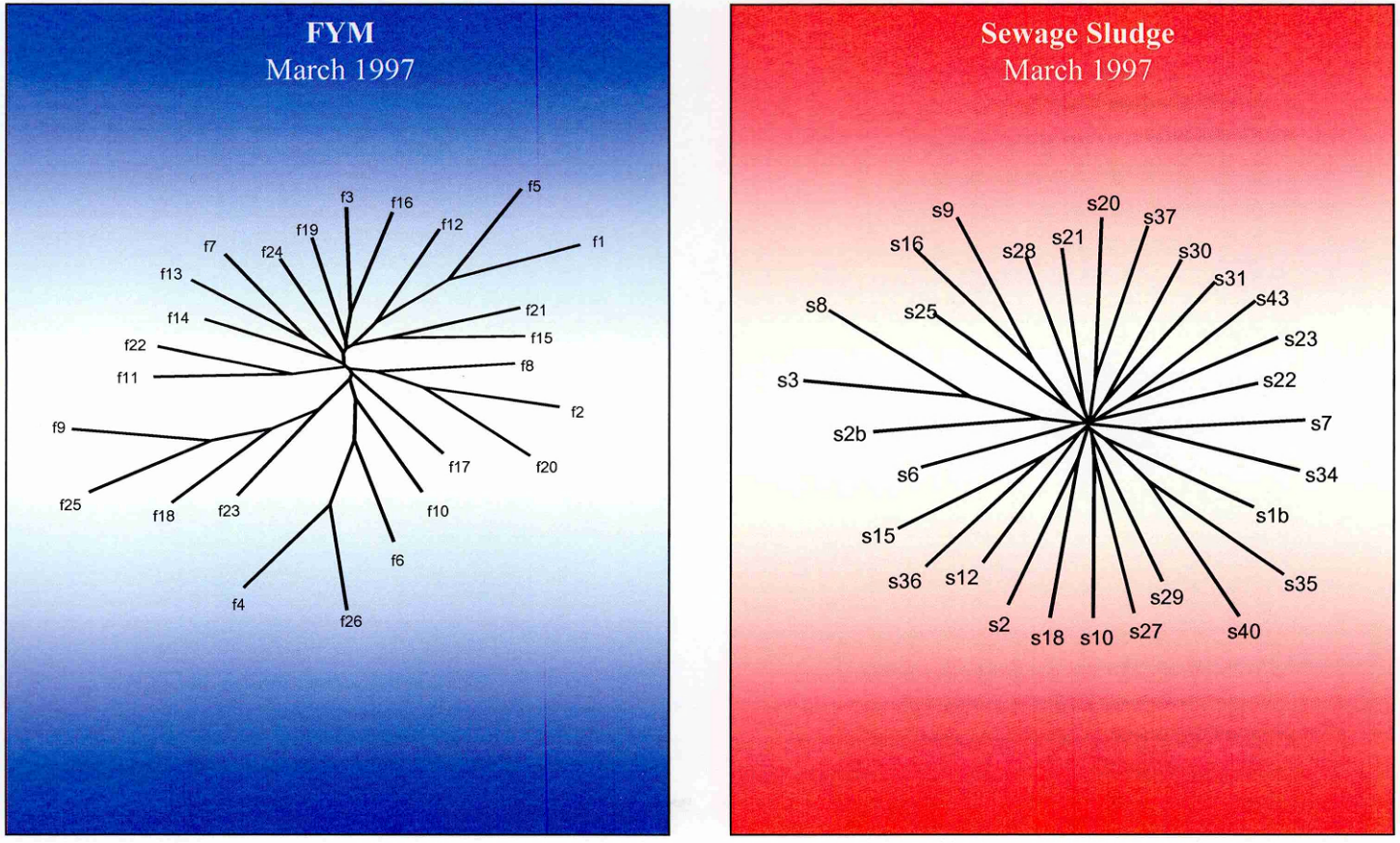


Figure 9.2b Population structure between plots (samples from 1998)

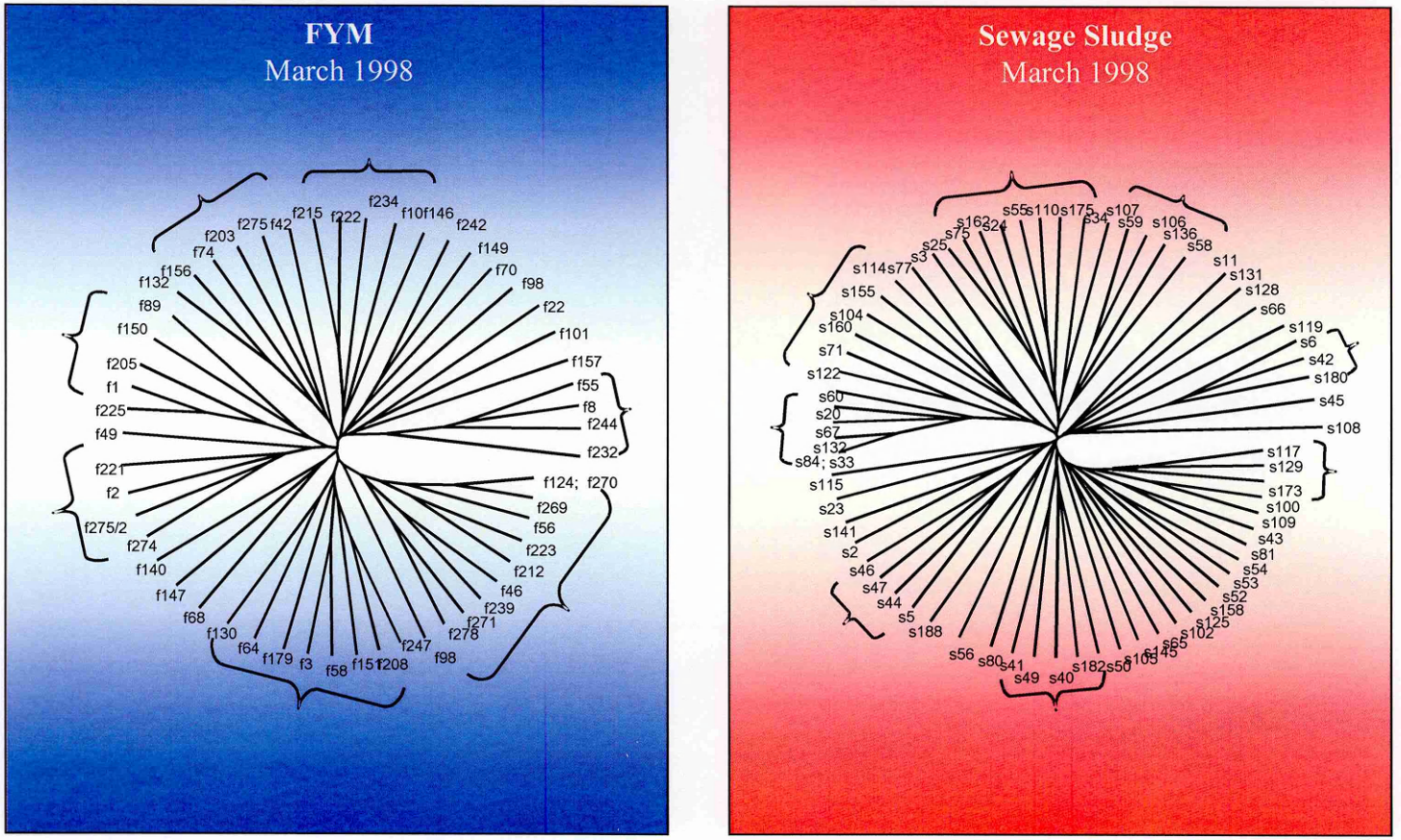
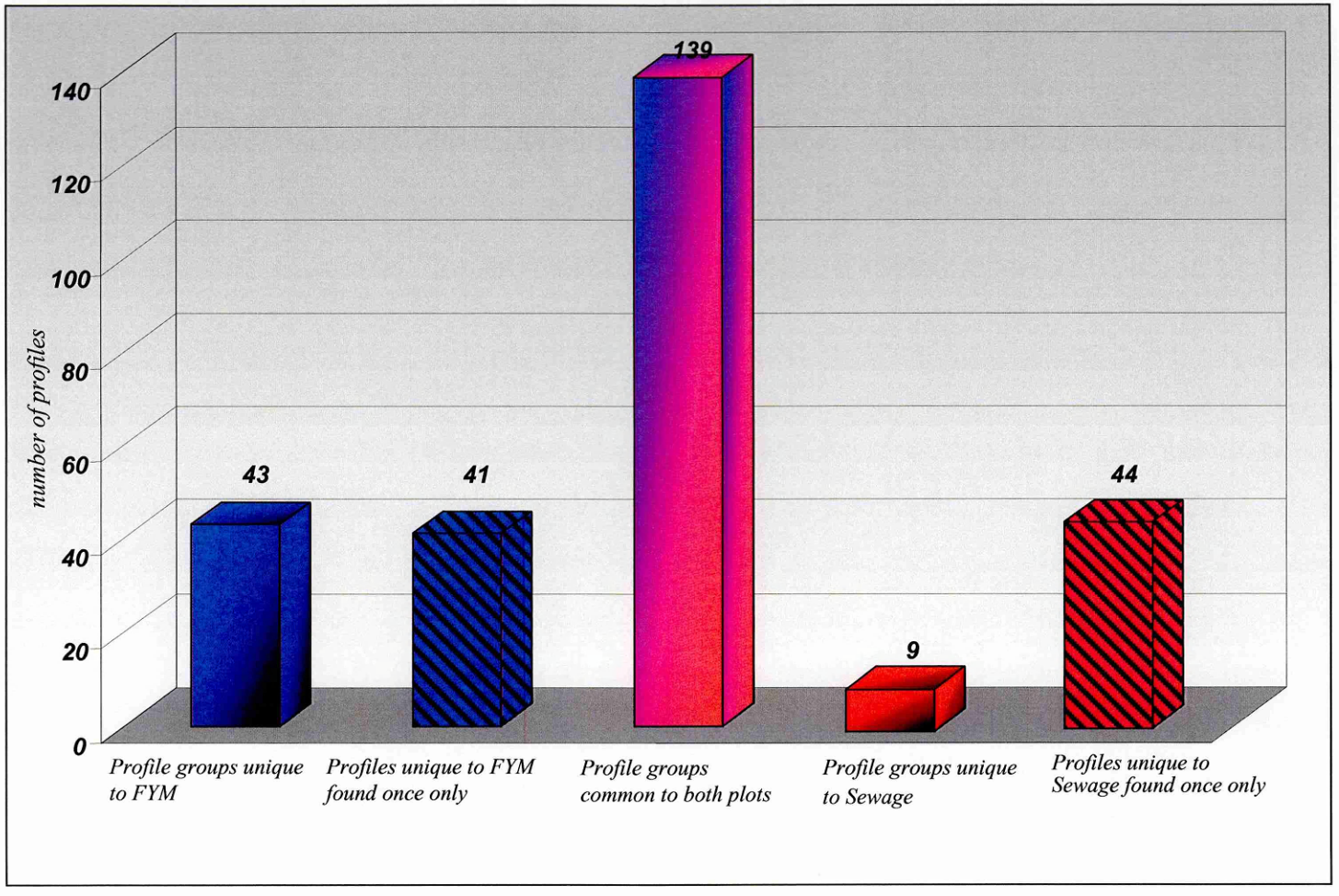


Figure 9.3. Differences (and commonalties) in distribution of profile types between plots



following March, the difference was not *clearly* seen, but there was nonetheless an indication towards the same trend (Figure 9.2a and 9.2b). The differences may be related to the heavy metals which have accumulated in the sewage plot, as well as to different fertiliser inputs. But, as the last fertiliser inputs were in the 1960s, so the organic matter will have degraded, with the metals remaining elevated indicating the net effect would be due to the heavy metal concentrations. Crucially, it indicates a possible difference in the population structure of the two groups, with **dominant groups** apparent in the **FYM plot**, unlike the **Sewage plot** (Figures 9.2 a & b). This is at least an intriguing contrast, given the frequently observed patterns of population asymmetries which often precede population crashes (section 1.5.1).

9.5.2.2 Conclusion

The results suggest that the pseudomonad community of the two Woburn plots had some differences in populations and somewhat different structures. This may be attributable to the presence of heavy metal contaminants, although the impacts are distorted by different organic matter inputs. Comparison of different numbers of isolates makes dendrogram formats difficult to interpret and it was not possible to produce a meaningful dendrogram to compare all the isolates together.

9.5.3 Phylogenetic structure

The normal way of expressing the phylogenetic relationship of isolates is by using a dendrogram with **hierarchical format** (Chapter 6), as it is helpful in expressing relative influences on the microbial populations. There are limitations in the use of this methodology for displaying data (Curnow *et al.*,1999); the major limitation in this instance is the actual size of the data.

The total fingerprint data was analysed using the Genstat Fingerprinting Programs. A similarity matrix (calculated using the DICE similarity coefficients) was obtained for the entire isolated pseudomonad population, but the overall size of the matrix made it impossible to obtain a phylogenic tree. The Phylip program (Chapter 6) was deployed in an attempt to obtain a dendrogram, but it was unable to accept the matrix 394 X 394 (i.e. the total number of profiles in the database after all analysis parameters were met).

The large matrix caused a fatal fault in the Phylip program. The problem proved intractable because of a segmentation error in the programming i.e. the complexity of the data's interaction caused the program to fall into a continuous loop, known as a segmentation fault (J. Felstein, personal communication, 2000).

9.5 3.1 Summary

With a small sub-set of profiles, a tree structure was inferred from the population, but with the expansion in the number of profiles, the tree structure was lost. The expansion of the soil isolate profile database, with both the increasing numbers and interactions, triggered the collapse in the dendrogram system. It can be assumed therefore that the dendrogram application has reached its limit, making it necessary to analyse and display the data in another format.

9.5.4 Similarity matrix

The database contains 394 successful isolate profiles, hence a matrix using the Dice equation (section 7.1) was calculated, giving a matrix of 394 X 394. The dendrogram format could not be used due to the size of the matrix (section 9.5.3), hence it was condensed into two presentational matrices, namely the "Parent Matrix" and the "Group Matrix" to allow easier interpretation of phylogenetic outputs.

9.5.4.1 Parent matrix

Table 9.8 includes all the isolates from the FYM and Sewage plots, i.e. the parent populations and it summarises the similarities between and within plots. The FYM plot groups have 44.1 % similarity to the 'parent' FYM plot (i.e. within the plot), while the Sewage plot population had 47% similarity within the parent group. This relates to the similarity between all the isolate profiles, within the plots, over all the sampling times. Whilst this result indicates slightly fewer types within the FYM population, relative to the total fingerprint database, the overall difference, at only 3%, is extremely slight. Therefore, it simply informs us that the difference within populations in the two plots is low.

Table 9.8 Similarity matrix between and within groups, using Dice

FYM Plot	44.1%	
Sewage Plot	**45.2%**	47.0%
	FYM Plot	**Sewage Plot**

The comparison of the two parent samples also reveals subtle (rather than significant) differences. Expressed as a percentage, this is a similarity level of 45.2%, which suggests that there is virtually as much in common, in terms of population profile, between the samples, as there is in terms of discernible difference. It is interesting to note that Chapter 3, which looked at Population Size and Composition Diversity, indicated the same result. So, the results from 1997, where small numbers of profiles were compared, were misleading, indicating more, rather than less, diversity in the sewage plot; and more, rather than less, structure in the FYM plot (O'Flaherty *et al.*, 2001). This is an important finding, as it highlights both the problems associated with interpreting ecological data and the danger of a single sampling approach.

9.5.4.2 Group matrix

Table 9.9 shows the percentage means of similarities between, and within, groups. It also displays the relationship between the FYM and Sewage sub-groups. It is also intended to display the relationship within the subgroups at different time intervals. The percentage variation within the FYM plot is between 34.4-75.8 %, while variation within the Sewage plot is between 41- 47.7%.

Table 9.9 Percentage means of similarities between and within groups, using Dice similarity matrix

FYM Mar 98	34.5							
FYM Mar 97[a]	48.1	75.8						
FYM Mar 97[b]	43.3	67.3	60.3					
FYM Oct 98	36.8	52.5	48.2	40.9				
Sewage Mar 98	39.3	58.7	53.5	43.4	47.9			
Sewage Mar 97[a]	41.3	63.1	57.1	45.4	50.7	53.5		
Sewage Mar 97[b]	43	66.1	59.6	48.2	51.9	56.3	60.8	
Sewage Oct 98	35.4	52	47.1	39.3	41.9	44.1	47.7	41
	FYM Mar 98	FYM Mar 97[a]	FYM Mar 97[b]	FYM Oct 98	Sewage Mar 98	Sewage Mar 97[a]	Sewage Mar 97[b]	Sewage Oct 98

[a] *and* [b] *= two isolation at one time point*

The most conspicuous factor from the tabulated data above is that the sampling for March 1997 seems to have resulted in a greater similarity within sites than was found at the other sampling times. A possible explanation for this is that the first data set was tested while all the systems were being developed. The development of the isolation and lysing procedures may have initially triggered a bias towards more easily cultured and lysed sub-populations, meaning that more recalcitrant species were lost. In contrast, greater similarities are observed within and between the two plots in March 1997, compared with other sampling points.

The results show that the populations in both plots have a lot in common at any given sampling point. However, the variation *intra-plots* at different sampling points is evident. It can reasonably be proposed that because of the high percentage of similarities between plots, *differences* between them are being camouflaged. Therefore in an attempt to highlight such possible differences, the simple matching program- the ERIC Matching Program- was employed, with the ability to discern differences between plots more readily.

9.5.5 Conclusions

The population diversity across (and within) both plots highlights a number of key considerations:

1. The effect of sampling regimes on the population profile appears to be highly significant in determining the types of sub-populations captured.
2. The variations between these sub-populations (i.e. genetic diversity) have a degree of consistency, namely around 45% similarity.
3. Importantly, therefore, levels of genetic diversity appear to remain proportionally the same, irrespective of the sampling regime.
4. The ERIC matching program may emphasise the differences between the plots.

9.6 ERIC matching program

The criteria for standardised parameters were used in the Microsoft Access Fingerprint Database Program, and profile data was then analysed using the **ERIC program** (section 9.1.4). Using this program, the distribution of isolate profiles in each plot, and their inter-relationships between plots, became apparent. The results are summarised in Figure. 9.3 and Table 9.10. The data outputs, shown in Figure 9.3, were compiled into 5 distinct groups, as follows:

Group 1: profiles common to both the FYM and sewage plot

The underlying diversity of the populations of both plots has many similarities. The predominant cluster contains bacterial groups common to both plots. ERIC-fingerprinting is clearly displaying what other diversity studies implied (Chapters 3 and 4).

Group 2: profiles common to the FYM plot which form into discrete groups

This includes common groups clustered with other isolates in the FYM plot, but lacking any similarity to Sewage isolates. This is an indication of a possible clonal population structure within the FYM isolates.

Group 3: profiles unique to FYM plot which do not form into discrete groups

This includes the large number of unique isolate profiles occurring in the FYM plot only once. They are not grouped with any isolates in the FYM or Sewage plots. This expresses the underlying richness of the system, detected using the multi-sampling approach.

Group 4: profiles common to the sewage plot which form into discrete groups

This group contains profiles which are grouped with other isolates within the Sewage population, but which have no similarity to the FYM isolates. This represents only a small proportion of the total profiles of the Sewage plot and is proportionally much smaller than in the FYM plot. There appears to be less population structure within the Sewage plot (i.e. the formation of groups within the plot) compared to the FYM plot.

Group 5: Profiles unique to Sewage Plot which do not form into discrete groups

This group displays the large number of unique isolate profiles found in the Sewage plot. They are not grouped with any other isolates in the Sewage plot, or the FYM plot. This group is proportionally much larger in the Sewage plot, indicating that there are more isolates in this plot which are unique and 'ungrouped', highlighting the relative absence of population structures compared with the FYM plot. This means that the sewage plot appears to have more *species richness* and less *evenness* than the FYM plot.

Table 9.10 Summary of grouping data showing the distribution of group types in each plot.

%	Sewage	FYM	Mixed
Unique	33	29	N/A
groups	7	30	63

9.6.1 Summary

These fingerprinting techniques illustrate two key findings:

1. The first relates to the **methodology deployed**: The diversity of the sampling regime has revealed a hugely diverse population structure. This would appear to validate the theory advanced by ecologists (Begon *et al.*,1986, Watkinson 1998, Hammond, 1995) that multi-sampling techniques pick up more unique species/profiles than a one-off population sampling.
2. The second highlights very significant levels of **bacterial redundancy** in the pseudomonad populations across the two plots (Table 9.10). Behind the prevalent clusters of related groups, there are a variety of unique strains. Based on the established functional patterns of pseudomonad communities, these unique populations possibly have the ability to carry out functions which are currently performed by the majority of the bacterial population. However, the possibility remains that not all isolates are related and that 'single' profiles may represent some unrelated genera, but happen to grow on PSA selective media.

9.7 Discussion

Comparison of the soil isolates from October 1998 showed that many shared either similar 16S rDNA genes (determined by ARDRA) or genomic DNA (determined by ERIC), or both, whether or not they amplified with the *Pseudomonas* 16S specific primers, indicating a degree of genetic relatedness. Therefore, despite this demonstration that a majority of isolates on PSA were *Pseudomonas*, it is still likely that some were **not** closely related to the genus, which might account for some of the large number of profiles occurring only once. The pseudomonad specific primers were tested against a number of different *Pseudomonas* sequences in the databases and the results showed (data not shown) that a number of species-classified as *Pseudomonads*- lacked the designated pseudomonad specific primer sites.

Therefore, although we can assume that some of the isolates which did not amplify with specific primers were not pseudomonads, it is unclear what percentage of isolates could belong to the section of clearly defined pseudomonads not having these specific sites.

In a previous (and much smaller) study on the diversity of PSA isolates from wheat under arable cultivation, only a limited number of groups was identified on the basis of their ERIC profiles (Nicholson *et al.,* 1998). The FYM and Sewage plots have not been cultivated since grass was sown in 1989, which might increase soil heterogeneity and hence bacterial diversity, both in the abundance of different groups and the number of isolates in each group.

UPGMA analysis of ERIC profiles from the first sampling indicated that the population in the FYM plot might be more structured or related (i.e. descended from fewer ancestral types) than the Sewage plot. This phylogenetic relationship can only be inferred due to the nature of the ERIC fingerprint. Analysis of greater numbers of isolates in subsequent years did not support this initial finding, with both dendrograms and DICE similarity coefficients showing no clear differences. One of the main findings of the research is the effect of a single sampling regime (O'Flaherty *et al.,* 2001), which has been shown, with subsequent samplings, to provide a skewed insight into the population in the two sites. Subsequent samplings (i.e. increased isolate numbers) did not conclusively support the initial findings.

Such an outcome has to allow for the possibility that there are in fact no meaningful differences and that *both* plots have diverse populations, with some isolates in common, and some unique, probably arising from the chance distribution of isolates in the soil at the start of the experiment, before the sites were divided into plots and treated with FYM or sewage sludge.

To summarise, there were fewer groups unique to the Sewage plot than to the FYM plot, although about half of all groups were found on both plots. Generally, the total number of different profiles identified increases with sample size, but this measure of '*richness*' was comparable in both plots. Moreover, the number of individuals in each group, or '*evenness*', was similar. In populations under moderate stress, previously-dominant groups may lose their advantage compared to the majority of individuals, resulting in an apparent increase in '*richness*' arising from the increases in '*evenness*'; conversely, if many groups become uncompetitive, these diversity measures may

decrease (Giller *et al.,* 1998). However, there are trends which indicate a slight increase in the diversity of the heterotrophic bacteria in the Sewage plot, compared with the FYM plot, but no firm conclusions can be drawn from the data presented in this thesis. Further research into the pseudomonad group would be required to discern the population dynamics of this group. The results suggest that further investigation into bacterial shifts in bacterial populations must be carefully considered and that multi-sampling is the strategy required to support a clear understanding of relationships in the soil environment

9.8 Conclusions

The *Pseudomonas* genus, its taxonomy and phylogenetic arrangement is often questioned (Anzai et al., 2000). This bacterial group contains a broad span of metabolic activities, with an inherent ability to survive complex and challenging environments. The metal resistance systems inherent or ubiquitous to members of the pseudomonad group are thought have arisen soon after life began in a world already polluted by volcanic activities and other geological sources (Bruce, 1997). This may explain the readiness of the pseudomonad group to survive the slow addition of metal contaminants over time. Or, more likely, the pseudomonads can be seen as a broad group of opportunist heterotrophics, which have evolved to exploit a wide variety of environmental niches. The criteria on which genetic diversity in soil bacterial populations is based are not as tightly defined as in some other groups of organisms. Hence, unlike laboratory phylogenetic studies, soil bacterial strains do not have to belong to a single cluster and may cluster in many different groups. The ERIC profiling allows for the concept of partial group membership, but it makes interpretation of population structures difficult and care must be taken when applying standard Phylogenetic logic.

If the knowledge obtained from this study is to be extended to other bacterial groups, it is nonetheless important to emphasise that we have, to date, only a partial understanding of soil population dynamics. Although this study has shown that subtle shifts in population structures occur under mild heavy metal stress, it equally highlights the limited knowledge we have of the underlying structure of the soil ecosystem.

Chapter Ten:
Conclusions and Implications

10.1 Introduction

This concluding chapter is designed to articulate the emerging hypothesis from the preceding environmental analysis. It highlights the key caveats and goes on to map out possible routes for further investigation. It will also suggest ways in which the findings can be used to investigate further, related environmental issues, by improving our understanding of the patterns governing microbial populations. It tackles these issues, using the following structure to draw out key conclusions. Firstly, there is an appraisal of the three different diversity assessment techniques (population size, catabolic and genetic diversity in studying soil environmental stress). Secondly, a hypothesis shaped by the output from this analysis and thirdly, future direction for the continuation of this work.

10. 2 Diversity assessment techniques: population size and composition

Population diversity provides a high level profile of heterotrophic communities, but its value is confined to a description of absolute numbers, rather than the changing composition of sub-populations. Population and composition diversity measurements for pseudomonads do not, by themselves, generate substantive conclusions, for two central reasons:

1) Pseudomonads are heterotrophic: Pseudomonads vary widely in their nutritional and physiological abilities, which means that population measures, in isolation, risk concealing much more significant changes, like the loss of key functions. The *Pseudomonas* genus has been re-evaluated many times (Anizie *et al.,* 2001, Fox *et al.,* 1992). The vagaries of working with these culturable micro-organisms are in contrast to the well defined *Rhizobium* populations. *Rhizobium* populations in metal contaminated sites (Hirsch *et al.,* 1993, 1996) can, due to their defined abilities (i.e. their ability to nodulate plants), be more effectively monitored. Single species with important (and understood) ecological functions act as sensitive 'early warning systems' and are not in any way prejudiced by the use of an otherwise cruder technique (i.e. assessment by

population size). The crashing of *Rhizobium* communities in metal contaminated plots is a classic illustration of the value of a population community measurement (Hirsch *et al.*, 1993, Giller *et al.*, 1998).

2) Population and composition measurements: The measures themselves provide only a general 'headcount' of heterotrophic populations, without any depth of information on *structural* consequences. The methods provide basic information on the size of the active communities, but do not on their own provide information on the bacterial community structure (Chapter 3). Hence it is possible to observe limited changes among the population while the occurrence of genetic re-configuration and changes in distribution of functional tasks remain undetected. They do not, in other words, clarify our understanding of the broader consequences of environmental degradation on soil microbial functions.

10.2.1 Summary

The differences in the populations with the three different samplings demonstrate the need for successive sampling. The multi-sampling strategic approach is designed to pick up multiple populations. This approach is advocated by environmental ecologists (Hawksworth 1995, Begon *et al.,* 1986) but is not an approach traditionally used in microbial ecology studies. Given traditional assumptions that multiple samplings increase the likelihood of incompatible results (to the point where populations are so unrelated that comparative research becomes unintelligible), this research output appears to be unexpectedly robust.

The result of the first sampling highlights the danger of a single-sample approach, showing the possibility of over-interpretation from a single sample. Although a trend has been identified in the data showing a difference between the FYM plot and the Sewage plot (and a possible increase in numbers in the Sewage plot), the results over time, on the bacterial population structure, imply that the relationship between the isolates is more complicated than the first sampling suggests. There is some level of consistency within plots, which points to divergence between plots.

10.2.2 Conclusions

The meaning in the measurements of pseudomonad population levels can only be properly understood by combining results from the population study with the application of genetic profiling of the pseudomonad population. These accordingly provide further insight into the size and the genetic composition of the pseudomonad populations.

10.3 Diversity assessment techniques: catabolic

Measures of catabolic diversity identify intriguing differences between FYM and Sewage plots, but provide only a simplistic or one-dimensional picture of changing patterns in microbial populations. Results from the catabolic diversity analysis indicate distinct differences between the two heterotrophic communities present in FYM and Sewage plots (Chapter 4), with no indication of what the difference may be. The established functional patterns of pseudomonad communities can be affected by bacterial functional redundancy. The concept of bacterial functional redundancy (Yin *et al.,* 2000) is if one species or group is lost, others have the ability to metabolise the substrate. This can be attributed to the fact that it essentially reflects the activity of a wide range of organisms which all have the ability to use the large number of carbon substrates present in the soil (Wenderoth and Reber 1999). The presence of smaller, isolated populations implies a latent ability to carry out functions currently performed by majority populations, which predominate in the main cluster. Therefore care must be taken in the interpretation of catabolic profiles.

To summarise, the catabolic study only reveals that there are differences in the heterotrophic populations in the two sites, but this does not provide any information on the bacterial population structure, or of the health of the soil. However, there are four central problems associated with outputs derived from this type of methodology:

1) The differences highlighted in catabolic diversity ought to be considered provisional until they can be corroborated in much greater depth. That means taking many more samples, over much longer time frames (as suggested by Grayston *et al.,* 2001). The methodology would benefit from additional ecologically relevant substrates to increase the amount, and therefore the quality, of data captured.

2) The apparent differences in catabolic diversity are also counter-intuitive, namely that bacteria in FYM have a more limited diversity than those in sewage plots. Taken by themselves, the results might suggest that the relationship between levels of soil pollution and catabolic diversity are completely inverse, because bacteria somehow become more resilient to contaminants over time than bacteria in pollution-free environments. But it is highly unlikely that microbial soil populations actually prosper the more that contaminants are added, so the relationship is likely to be more complex than the results by themselves imply (Wenderoth and Reber, 1999). My hypothesis is that as the dominance of the best adapted ecotypes is eroded due to metal contamination, so a more diverse population of ecotypes that was not previously visible, is revealed. Only the apparent number of species increases, not the actual number. There seems to be an increase in the number of species not an actual increase in numbers (section 4.3.1).

3) A limitation of this technique is its failure to profile a broader picture of shifts in microbial populations. While catabolic diversity can identify shifts in behavioural characteristics of heterotrophic populations, it can do nothing to *attribute* those shifts to a bacterial group or shifts in soil properties (Degens *et al.,* 2001). Bacterial population structure cannot be observed using this technique in isolation; indeed, changing patterns in microbial diversity may not necessarily equate to changes in functional diversity. The question needs to be settled by more in-depth study. In the meantime, total bacterial diversity can be considered as a surrogate measure of biological diversity, which needs to be corroborated with other studies.

4) Catabolic measures lack refinement. The use of BIOLOG GN1™ for monitoring effects in moderately contaminated soils, typical of the situations relevant to EU sewage sludge limits (Lawlor *et al.,* 1998, O'Flaherty *et al.,* 2001), may not be as useful an indicator as it can be in more polluted soils. Indeed, the BIOLOG GN1™ method is effective at detecting significant differences in microbial populations when there is an 'acute' (i.e. substantial and immediate) stress effect on heterotrophic populations, e.g. a combination of low soil pH and heavy metal-contamination (Knight *et al.,* 1989) and pollutant nitrogen deposition (Johnson *et al.,* 1998). Also the selection of the BIOLOG GN1™ plates themselves reveal degrees of soil stress more efficiently than

EcoPlates™. BIOLOG GN1™ are a superior methodology than EcoPlates™ by virtue of the larger number of substrates provided (section 4.3.2).

10.3.1 Conclusions

The long term *'chronic'* stress effect (i.e. sustained, incremental pollution) of past sewage applications on populations is more complex than studying acute stress on soils. Understanding the effect on bacterial population structures needs more detailed study than catabolic measures alone can provide. Catabolic diversity therefore has utility only as a *first* line of investigation, acting as an essentially observational, rather than an analytical tool.

10. 4 Diversity assessment techniques: genetic

Genetic assessment complements our understanding of pseudomonad populations by revealing selective structural differences between plots. The findings in the genetic study complement findings of both catabolic and population diversity studies, but allow for a more in-depth understanding of complex bacterial shifts within stressed and non-stressed soils.

10.4.1 Pseudomonad population structure

In the main, genetic diversity identifies similar 'core' characteristics between plots, but despite strong similarities in structure and population, there are marked differences in 'outlying' populations (Barnett and Lewis, 1983), which are smaller and characterised by different interrelationships: greater population structure in FYM; more unique fingerprints in sewage (Chapter 9). The phylogenetic analysis of the bacterial isolates due to the nature of the ERIC profile information only allows us to infer phylogenetic relationships. The initial phylogenetic studies produced a dendrogram (section 9.5.2), which indicted that the populations within the FYM plot were more clonal and reflected a defined bacterial system which might display competitive exclusion (Lynch and Hobbies 1988). On the other hand, the Sewage Plot appeared to have a less clonal structure, indicating initial signs of a change of this structure within the soil system.

Unlike lab-based phylogenetic studies based on genetically defined bacterial sequences, the unpredictable behaviour of soil isolates and their inferred relationships impede the

clarity of soil-based genetic profiling. Soil bacterial strains do not necessarily belong to a single cluster and may instead cluster in multiple groups, where they are characterised by partial, or 'shared', membership (Fox *et al.,* 1992). Interpretation of population structures is difficult and care must be taken when applying standard phylogenetic laws and logic as the analysis present in this thesis is not DNA sequence based. However, the non-phylogenetic grouping of profile data (Figure 9.3) shows that the FYM plot tends to be more clonal than the Sewage plot. The number of isolates need to be increased to verify the initial indications. However the research here has shown the difficulty of working with such a genetically diverse group as pseudomonads.

Therefore, it can be concluded that the pseudomonad group may not be suitable for application of a general biological indicator to soil health. The development of strategies and techniques presented in this thesis can now be easily transferred to another target group of soil organisms in search for a biological indicator to soil health.

10.4.2 Summary

Genetic analysis of the soil isolates, via either the genomic DNA (determined by ERIC), or 16S rDNA genes (determined by ARDRA), or both, indicates a degree of genetic relatedness. Despite this demonstration that a majority of isolates on the selective media (PSA) were *Pseudomonas*, it is likely that some bacterial isolates may not be as closely related to the genus, which may account for the large number of unique profiles. Generally, the total number of different profiles identified increases with sample size, but this measure of "richness" was comparable in both plots. Moreover, the number of individuals in each group, or "evenness" was similar. In populations under moderate stress, previously-dominant groups may lose their advantage compared to the majority of individuals, resulting in an apparent increase in "richness" arising from the increases in "evenness"; conversely, if many groups become uncompetitive, these diversity measures may decrease (Giller *et al.,* 1998).

10.4.3 Conclusions

Trends have been identified in the two plots but further investigations would be required to provide conclusive evidence of an increase in diversity of the heterotrophic bacteria in the Sewage plot, compared with the FYM plot. The FYM and Sewage plots have not

been cultivated since grass was sown in 1989, which might increase soil heterogeneity and hence the apparent bacterial diversity, both in the abundance of different groups and number of isolates in each group.

10.5 Complementary to catabolic & population studies?

Catabolic and population diversity are consistent and point to some effects on the distribution of the pseudomonad types in each plot. This suggests the effects of heavy metals are subtle on populations studied and demonstrates the ability of the system to show subtle changes in the genetic diversity .

There are **four** conceivable hypotheses emerging from this thesis, which may explain the difficulties with uncovering the differences between the two soil plots. Intriguingly, none of these hypotheses is mutually exclusive and it is also hard to draw absolutely firm conclusions about the likely validity of any one of them. Nevertheless, some are more probable than others.

(1) Functional ability: It assumes that detectable differences between the two plots can be accounted for by differences in the functional *ability* of bacterial populations. This was highlighted by the analysis of catabolic and genetic diversity. It assumes, for example, that bacterial strains in the FYM plot have fewer capabilities that they can deploy to utilise carbon substrates, relative to bacterial populations in the Sewage plot, (i.e. an increase in the number of different species).

(2) Relative volumes: It assumes differences in absolute numbers of total bacterial sub-populations relative to their ability to utilise certain carbon substrates. In the FYM plot, for example, there could be far fewer bacterial strains able to target specific substrates, whereas in the Sewage plot there may be many more (i.e. the increased catabolic ability of strains).

(3) Population structure: It assumes that the discernibly different structures between the FYM and Sewage plots accounts for many of the differences. As the results indicated earlier, this was a predictable, statistically significant difference highlighted using genetic diversity modelling techniques. The overall species richness of each plot

may be the same, but it is masked in the more structured FYM population, by a few dominant types.

(4) Structural bias in methodology: One possibility that needs to be acknowledged, however, is the imperfections of the methodology overall. While differences have been implied, there are a number of imperfections in the approach and data analysis, which may have led to a simple overstatement of the case.

10.5.1 Overview

The results presented do not point unambiguously to one hypothesis in particular. Further evidence is needed on the ecosystem before any firm conclusion can be drawn on these hypotheses. It may be possible that one (or all of the above) scenarios are responsible in some way for the population shifts. Therefore no one hypotheses will be selected as the most likely, until further investigations are completed. However, the model of stress relationship and population shifts presented in Chapter 1 will be addressed.

10.6 Model of stress on population structure

This model proposed in Chapter 1 is that increasing levels of pollution have a cumulatively adverse affect on soil population diversity, assuming that the metal levels between the plots are the only differences. As those levels of stress increase, populations enter a phase of 'post' competitive exclusion (i.e. new emergent populations, once the prevailing order of species dominance is disrupted) followed by extinction, a pattern which can be profiled in a distinctive bell-shaped curve. The effect on sludge-treated soils of moderate stress, in this case heavy metal contamination, parallels this ecological theory.

The principal findings from this thesis were not conclusive, but the trends observed could be reflected in Austin's hypothesis, showing how the effect of increasing stress affects diversity. However further research will be needed to draw firm conclusions as the results do not clearly reflect an increase in diversity. The following three reasons currently prevent us from drawing firm conclusions from the thesis:

1) Excessive amounts of pollution are undoubtedly damaging to population levels. Pseudomonad populations, possessing a broad collection of metabolic activities, have an inherent ability to survive and to adapt challenging environments. In accounting for this resilience, it is important to remember that metal-resistant properties are thought to have evolved soon after life began, in a world already polluted by volcanic and other geological emissions (Bruce, 1988). Nevertheless, the ability of pseudomonads to survive the slow, incremental accretion of metal contaminants, would not necessarily apply where there are much heavier levels of pollution

2) The sheer breadth of the pseudomonad genus may conceal specific population loss. In contrast to the single species/single function monitoring, a large range of bacteria from the pseudomonad group can show changes in diversity at a global level. As with *Rhizobium*, single species with specific ecological functions are more likely to be sensitive, acting as 'sentinels' of change, but because of the breadth of the pseudomonad populations being profiled, some sensitive types may elude detection (i.e. loss of these sensitive pseudomonads may go undetected). This could be compatible with measured increases in overall diversity.

3) Plant communities' interactions may not be imitated by bacterial communities. An obvious and important point to highlight, however, is that not all microorganisms will behave like the plant kingdom. For a start, microorganisms have a high level of redundancy in their population structure, which is not reflected to the same level in the plant kingdom. Therefore a theory can be postulated that there are numerous bell shape curves reflecting different bacterial groups, reflecting a very complex picture of soil bacterial communities. But the purpose in drawing a cross-species analogy such as between plant and soil populations is to corroborate the mounting evidence which points to soil population increase, as indicated here, followed by their demise (in pseudomonad populations and in specific populations like rhizobia). Further research will be able to test the strength of these relationships and the patterns they exhibit.

10.7 Summary conclusions

Despite evidence of bacterial redundancy, any interim position on the effects of metal contamination should err on the side of caution. What is understood about pseudomonad

dynamics in soil is limited, and what we know about the profusion of remaining bacterial groups, even less. Indeed, as Margules argues, genetic diversity in soil bacteria needs to be conserved at a species richness level to ensure that *'the whole suite of species'* is conserved (Margules *et al.,* 1982), therefore advocating that sensitive bacteria need to be used as sentinels of change, and that the increasing levels of pollution have an acute effect on soil microbial diversity. More research is required to understand the effects of pollution on the population structure of the pseudomonad group, but also the cumulative results have shown a subtle differences, even if at present our understanding of the results are limited.

10.8 Future directions

This thesis has endeavoured to extend our understanding of bacterial shifts in heavy metal contaminated soils. It has taken a distinctive multi-disciplinary methodology to the study of complex interactions in the soil environment and applied it to a broad group of bacteria, from which we can draw a more rounded view.

Nonetheless, there is still no extensive perspective on soil population dynamics, compounded by our limited knowledge of other bacterial groups in soil. Although this study has detected the subtle shifts in population structure under mild abiotic stress, it necessarily leaves unanswered a wholes series of additional questions about the wider structure of the soil ecosystem. Overall, therefore, it re-enforces the value of comprehensive environmental analysis to act as a platform for a deeper understanding of microbial communities. Clearly one central benefit of that will be the determination of a commonly understood threshold of what constitutes tolerable, and what deleterious, levels of contaminants.

Where this thesis leaves off, unsurprisingly, is on the threshold of a series of open-ended questions, some of which we consider in turn below, and all of which should form the basis of future research.

10.8.1. How can we use extended methodologies to profile soil populations with greater accuracy?

Despite the prevalence of data which indicates divergent population profiles between plots, some allowance should be made for distortions caused by other variables, in particular different organic matter inputs. Quantifying this qualification is impossible without separate and extended study, but this consideration should inform subsequent research.

10.8.2. How can we deepen our current understanding of functional diversity?

We would also propose a significant and innovative approach to structure further investigation, namely mapping bacterial shifts over time, by tracking the rhizosphere of plants over growing seasons. There are two virtues to this approach. Firstly, such a method would allow continuous monitoring of bacterial populations, allowing us to differentiate shifts, which are attributable to environmental factors (whether biotic or abiotic) and those governed by 'natural' changes, for example in the nature and extent of plant exudate. Secondly, this approach allows us to aggregate traditional physical soil science measurements with biological measurements favoured by microbial ecologists. Crucially, this would allow us to obtain deeper insights into functional diversity, for example by attributing shifts in the physical state of soils, perhaps through increased nutrient cycling, alongside proportionate increases in pseudomonad populations.

10.9 Overall Conclusions

The purpose of the study was to investigate the effects of chronic metal pollution due to long-term sewage sludge application, on the population size, diversity and catabolic activity of soil pseudomonads. We were especially interested in the possibility that population diversity might be affected, and initial results indicated a possible impact with a small reduction in population size, a divergence in catabolic activity, and a change in population structure compared to a FYM-treated control plot (compared with the Sewage treated plot).

We have developed a methodology which allowed us to compare a large number of different PCR fingerprints and enabled the comparison of a relatively large number of isolates to be completed. However, the extended study over a three-year period did not clearly support the initial findings, but showed a greater effect between sampling times

than plot differences. The conclusion from the catabolic study and population composition study which reflects the activity of fast-growing heterotrophic bacteria (including many pseudomonads), were inconclusive and only alluded at the subtle differences between the plots. Overall, there was no clear difference in the main group of *Pseudomonas*-related heterotroph population structure, with similar diversity in both plots. The only differences between the plots concentrated in the 'outlying populations', which was highlighted in the genetic diversity study.

Thus, in contrast to rhizobia research, soil pseudomonads have not been demonstrated to be sensitive indicators of relatively low levels of heavy metal pollution. Indeed, the use of the pseudomonad groups as a biological indicator of soil quality did not achieve sufficiently clear results for us to propose it as a biological indicator. *Rhizobium*, in contrast, are proven to be a sensitive indicator of soil quality and health in the assessment of the effect of chronic metal pollution due to long-term sewage sludge application.

There remains a need for a general biological indicator of soil quality, but this work does not support the application of the pseudomonad group as the most suitable bacterial group. However, the systems developed, and the strategies highlighted in this thesis, can now jointly be used in the future to test other possible bacterial groups which may prove suitable as a general biological indicator.

It may be that the pseudomonads are not a useful indicator group for low level chronic stress and that sensitive indicator species are more useful. Diversity *per se* may not be a sensitive indicator with a complex group such as the pseudomonads. Predictably, this research is open-ended, leaving a distinctive trail of issues for supplementary research to pursue. In essence, this work has prized open a small crack in the door; it is for others to try to open it fully.

References

Anzai, Y., Kim, H., Park, J-Y., Wakabayashi H., & Oyaizu, H. (2000). Phylogenetic affiliation of the pseudomonads based on 16S rRNA sequence. International Journal of Systematic and Evolutionary Microbiology **50:** 1563-1589.

Akkermans, A.D.L., van Elsas, J.D., and de Bruijn, F.J. (1995). Molecular Microbial Ecology Manual. Kluwer Academic Publishers, Dordrecht, The Netherlands, Vol 1: 1-488.

Alef, K. & Nannipieri, P. (Eds) (1996). Methods in Applied Soil Microbiology and Biochemistry. Academic Press, London

Allsopp D., R.R. Colwell and D.L. Hawhsworth (1995). Microbial Diversity and Ecosystem Function. Published by CAB International.

Atlas R.M and Bartha M.J.(1998). Molecular Ecology Fundamentals and Applications Appendix 1: Statistics in Microbial Ecology. Published by CAB International

Austin M.P. and Smith T.M (1989). A new model for the continuum concept. Vegetation **83:** 35-47

Baath E., M.Diaz-Ravina, A. Frostegard and C.D. Campbell (1998). Effect of metal-rich sludge amendments on the soil microbial community. Applied and Environmental Microbiology **64:** 238-245.

Babich H. and Stotzky G. (1980). Environmental factors that influence the toxicity of heavy metal and gaseous pollutants to micro-organism. CRC Critical Reviews in Microbiology **8:** 99-145.

Barnett, V. & Lewis, T. (1993). Outliers in statistical data. 2nd. Published by Wiley, Chichester, U.K.

Begon M., Harper J.L., and C.R. Townsend (1986). Ecology Individuals population and communities, Chap 4 Life and Death in unitary and modular organisms. Published by Blackwell Scientific Publications.

Bengtsson J. (1998). Which species? What kind of diversity? Which ecosystem function? Some problems in studies of relations between biodiversity and ecosystem function. Applied Soil Ecology **10:** 191-199.

Bergey, D. H. & Kreig, N. R. Ed (1993). Bergey's Manual of systematic bacteriology. Vol. 1. Published by Williams & Wilkins, Baltimore, U.S.A.

Bernard M., Humbert O, Camara M., Guenzi E., Walker J., Mitchell T., Andrew P., Prudhomme M., Alloing G., Hakenbeck R., Morrision D.A., Boulnois G.J. and Claverys J.P. (1992). A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Research **20:** 3479-3483.

Biolog (1993). Instructions for the use of the Biolog GP and GN Microplates. Biolog Inc., Hayward, Calif, U.S.A.

Bochner B.R. (1989). Sleuthing out bacterial identities. Nature **339:**157-158

Bochner B.R. (1989). Breathprints at the microbial level. American Society for Microbiology News. 55: 536-539.

de Bruijn F. J., M. Schneider, U. Rossbach and F.J. Louws (1995). Automated Fluorescent and Conventional Rep-PCR Genomic Fingerprinting and Multiplex PCR to Classify Bacteria and Track Genes. Proceedings of the 7th International Symposium on Microbial Ecology Brazil 1995:1-8.

de Bruijn F. J. (1992). Use of Repetitive (Repetitive Extragenic Palindromic and Enterobacterial Repetitive Intergeneric Consensus) Sequences and the polymerase Chain Reaction to Fingerprint the genomes of *Rhizobium meliloti* Isolates and other soil bacteria. Applied Environmental Microbiology, **58:**2189-2187.

Brooks P.C. and McGrath S.P (1994). Effects of metals toxicity on the size of the soil microbial biomass. Journal of Soil Science **35:** 341-346.

Bruce K. D. (1997). Analysis of mer gene subclasses within bacterial communities in soils and sediments resolved by fluorescent-PCR-restriction fragment lenght polymorphism profiling. Applied and Environmental Microbiology **63:**4914-4919.

Burrows P.R., P.D. Halford and K. Evans (1996). Estimation of genomic variation between British populations of the potato cyst nematode *Globodera pallida* using RAPD-PCR. Molecular Ecology **5:** 697-701.

Campbell, C. D., Grayston, S. J., & Hirst, D. J. (1997). Use of rhizosphere carbon sources in sole carbon source tests to discriminate soil microbial communities. Journal of Microbiological Methods. **30:** 33-41.

Chander K. and Brooks P.R. (1992). Effects of heavy metals from past application on microbial biomass and organic matter accumulation in a sandy loam U.K. soil. Soil Biology and Biochemistry. **23:** 927-932.

Chen H.K, Liu X.C And Patel D.J (1993). Solution Structure of a Quinomycin Bisintercalator-DNA Complex. J Cell Biochem **10:** 261-261.

C.E.C. (1986). On the protection of the environment, and in particular of the soil, when sewage sludge is used in agriculture., Official Journal of the European Communities. 86/278/EEC.

Curnow R.N. (1998). Estimating Genetic Similarities within and between Populations. Journal of Agricultural, Biological and Environmental Statistics. **3**: 347-358

Davis K.G., F.A.M. De Leij and Kerry B.R. (1991). Microbial agents for the biological control of plant-parasitic nematodes in tropical agriculture. Tropical Pest Management **37:** 303-320.

Degens B.P. and Harris J.A. (1997). Development of a physiological approach to measuring the metabolic diversity of soil microbial communities. Soil Biology and Biochemistry **29:** 1309-1320

Degens B.P., L.A. Schipper, G.P. Sparling and L.C. Duncan (2001). Is the microbial community in a soil with reduced catabolic diversity less resistant to stress and disturbance. Soil Biology and Biochemistry **33:** 1143-1153.

Digby, P. G. N. and Kempton, R. A. (1986). Multivariate analysis of ecological communities. London, Chapman and Hall.

D.o.E. (1996). Code of Practice for Agricultural Use of Sewage Sludge. London, D.o.E., London.

Felstein J. (1996). PHYLIP V3.57 distributed by author. Dept. of Genetics, University of Washington, Seattle.

Felstein J. (1994). PHYLIP V3.57 (phylogeny inference package). Distributed by author. Dept. of Genetics, University of Washington, Seattle.

Fox G.E., Wisotzkey J.D. and J.R. Jurtshuk (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int. Journ. Syst. Bacteriol. **42:** 166-170

Garland J. (1996). Analytical approaches to the characterisation of samples of microbial communities using patterns of potential c source utilisation. Soil Biol Biochem **28:** 213-221.

Garland J. L (1997). Analysis and interpretation of community-level physiological profiles in microbial ecology. FEMS Microbiology Ecology **24:** 289-300.

Garland J.L and Mills A.L. (1991). Classification and characterisation of heterotrophotic microbial communities on the basis of patterns of community level sole carbon source utilisation. Applied and Environmental Microbiology **57:** 2351- 2359.

Genescan 672 User Manual (1993). Applied Biosystems, Inc., Perkin-Elmer Corporation.

Genstat Manual, (1988). Genstat 5 Release 3, reference manual, Clarendon, Oxford, UK.

Genstat Newsletter (1988), *22, 7-22.*

Gilbert G.S., M.K. Clayton J. Handelsman, and J.L. Parke (1996). Use of cluster and disriminant analyses to compare rhizosphere bacterial communities following biological perturbation. FEMS Microb. Ecology **32:**123-147.

Giller K.E., E. Witter and S.P. McGrath (1998). Toxicity of heavy metals to microorganisms and microbial processes in agricultural soils: A review. Soil Biology Biochemistry **30:** 1389-1414.

Goto M.(1986). Fundamentals of Bacterial Plant Pathology. Published by Academic Press Inc.

Grayston SJ, Wang SQ, Campbell CD, Edwards AC (1998). Selective influence of plant species on microbial diversity in the rhizosphere. Soil Biology & Biochemistry **30 :** 369-378

Grayston SJ, G.S. Griffith, J.L. Mawdsley, Campbell CD and R.D. Bardgett (2001) Accounting for variability in soil microbial communities of temperate uplands ecosystems. Soil Biology & Biochemistry **33:** 533-551.

Grothues D. and B. Tummler (1991). New approaches in genome analysis by pulsed-field gel electrophoresis: application to the analysis of Pseudomonas species. Molecular Microbiology **5:** 2763-2776.

Hammond P. M. (1992). Systematics and diversity: species inventory. In: Groombridge, B. (ed) Global Biodiversity: Status of the Earth's Living Resources. A report compiled by the World Conservation Monitoring Centre. Chapman & Hall, London, pp17-39

Hammond P.M. (1995). Practical approaches to the estimation of the extent of biodiversity in species groups From Biodiversity, Measurement and estimation edited Hawksworth Published by The Alden Press Chapman and Hall, U.K.

Hawksworth, D.L. (1991). The fungal dimension of biodiversity: magnitude, significance, and conservation. Mycological Research **95:** 641-655.

Hornby D., Bateman Gl, Payne Rw, Brown Me, Henden Dr and R. Campbell (1993). Field-Tests of Bacteria and Soil-Applied fungicides as Control Agents for Take-All in Winter-Wheat. Annals of Applied Biology **122:** 253-270

Hirsch P.R. , Jones M.J., , McGrath S.P. and G.E. Giller (1993). Heavy metal from past applications of sewage sludge decrease the genetic diversity of *Rhizobium leguminosarum biovar trifolii* populations. Soil Biology and Biochemsity **25:** 1485-1490

Hirsch P.R. (1996). Population dynamics of indigenous and genetically modified rhizobia in the field. New Phytol **133:**159-171.

Insam H. and A. Rangger (1997). Microbial Communities: Functional versus Structural Approaches. Published by Springer-Verlag Berlin.

Johnson D., Leake J.R., J.A. Lee and C.D. Campbell (1998). Changes in soil biomass and microbial activities in response to 7 years simulated pollutant nitrogen deposition on a heathland and two grasslands. Environmental Pollution **103:** 239-250.

Kempton, R.A. (1979). The structure of species abundance and measurement of diversity, Biometrics **35**, 307–321.

Kerry B.R. and J.M. Bourne (1996). The importance of rhizosphere interactions in the biological control of plant parasitic- A case study using Verticillium chlamydosporium. Pesticide Science **47:** 69-75.

Kloepper J.W., J. Leong, M Teintze and M.Schroth (1980). Pseudomonas siderophores: A mechanism Explaining disease-suppressive soils. Current Microbiology **4:** 317-320.

Killham K. (1985). A physiological determination of environmental stress on the activity of microbial biomass. Environmental Pollution **38:** 283-294.

Knight B.P. , McGrath S.P. and A.M. Chaudri (1997). Biomass carbon measurements and substrate utilization patterns of microbial populations from soils amended with cadmium, copper and zinc. Applied and Environmental Microbiology **63:** 39-43.

Laguerre G., M.Allard, F Revoy and N. Amarger. (1994). Rapid identification of rhizobia by restriction fragment length polymorphism analysis of PCR-Amplified 16S rRNA genes. Applied and Environmental Microbiology **60:** 56-63

Laguerre G., L. Rigottier-Gois and P. LeManceau (1994). Fluorescent Pseudomonas species categorised by using Polymerse Chain Reaction(PCR) and restriction fragment analysis of 16S rDNA. Molecular Ecology **3:** 479-487.

Lawlor, K., Knight, B.P., Barbosa-Jefferson, V.L., Lane, P.W., Lilley, A.K., Paton, G.I., McGrath, S.P., O'Flaherty, S.M., and Hirsch, P.R. (2000). Comparison of methods to investigate microbial populations in soils under different agricultural management. FEMS Microbiology Ecology. **33:** 129-137.

Lynch J.M. and J.E. Hobbie (1988). Micro-organisms in Action: Concepts and Applications in Microbial Ecology. Published by Blackwell Scientific publications.

Lynch J.M. (1986). Soil Biotechnology: Chapter 10 Epilogue: Treatment and Management. Published by Blackwell Scientific publications.

MAFF SP 0120 (1999). Impact of heavy metal on soil quality with respect to microbiological activity and production of crops. SP 0120, MAFF, London.

MAFF /DoE (1993). Review of the rules for sewage sludge application to agricultural land: Soil fertility aspects of potentially toxic elements. Report of the Independent Scientific Committee, PB 1561, MAFF publications, London.

Margules C.R., Higgs A.J. and R.W. Rafe (1982). Modern biogeographic theory: are there any lessons for nature reserve design. Biological Conservation **24:** 115-128.

Magurran A.E. (1988).Ecological diversity and its measurement. Published by University Press, Cambridge.

Manly, B.F.J. (1994). Multivariate Statistical Methods: a primer. 2nd edition. Chapman and Hall, London, U.K.

Milliar M., A. Wlodawer, E. Appella, J.L. Sussman (1987). Crystallization of a DNA Duplex 15-Mer Containing Unpaired Bases - D(CGCGAAATTTACGCG). Journal of Molecular Biology. **195:** 967

McGrath S P & C H Cunliffe (1985). A Simplified Method for the Extraction of the metals Fe, Zn, Cu, Ni, Cd, Pb, Cr, Co and Mn from Soils and Sewage Sludges. Journal of Science Food and Agriculture. **36** 794-798.

McGrath S.P. (1987). Long-term studies of metal transfers following application of sewage sludge. Special publication number 6 of the British Ecology Society 'Pollutant Transport and Fate in ecosystem edited P.J. Coughtrey, M.H Martin and M.H. Unsworth. Published Blackwell Scientific publications.

McGrath S.P. (1994). Effects of heavy metals from Sewage Sludge on Soil Microbes. Chapter 6 from Toxic metals in Soil-Plant Systems. Edited by S.M. Ross, John Wiley & Sons Ltd.

McGrath, S. P. (2001). How organisms live with heavy metals in the environment (Fact Sheet on Environmental Risk Assessment 2). 4pp. International Council on Metals and the Environment (ICME), Ottawa.

Miles A.A. and Mirsa S.S. (1949). The estimation of the bactericidal power of blood. Journal of. Hygiene. **38:** 732-749.

Murray M.G., and W.F. Thompson (1980). Rapid isolation of high molecular weight plant DNA. Nucleic Acid Research **8:** 4321-4325.

Nei M. (1987). Molecular Phylogeny and Human-Evolution. Journal of. Genetic. **61:** 633-633.

Nei M. and W.H. Li (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76: 5269-5273.

Nicholson P.S. and P.R. Hirsch (1998). The effects of pesticides on the diversity of culturable soil bacteria. Journal of Applied Microbiology **84:** 551-558.

Niemann, P. (1991). Breeding for herbicide resistance: benefits, risks and perspectives Plant research and development. **34**: 14-33.

O'Donnell A.G. , M. Goodfellow and D.L. Hawksworth (1994). Theoretical and practerical aspects of the quantification of biodiversity among microorganisms. Phil. Trans. R. Soc. Lond. **345:** 65-73.

O'Faherty S.M, Degens B., Murphy D.V., and Goulding K.W.S (1998). Microbial functional Diversity in agricultural soils. The 5th Congress for European Society for Agronomy.

O'Flaherty, S. M. & Coleman, K. W. (1999). Database processing of PCR fingerprinting of bacterial isolates from agricultural soils (Microsoft Access database). IARC Rothamsted, Harpenden, U.K. (unpublished, request reprint No: 16540)

O'Flaherty, S., McGrath, S.P. & Hirsch, P.R. (2001). Comparison of phenotypic, functional and genetic diversity of bacterial communities in soils. In *Sustainable Management Of Soil Organic Matter* (eds R. M. Rees, B. C. Ball, C. D. Campbell & C. A. Watson) 370-376. CAB International, Wallingford.

Pace N.R., D.A. Stahl, D.J. Lane and G.J. Olsen (1986). The analysis of natural Microbial populations by ribosomal RNA sequences. Advances in Microbial Ecology. 9:1-55. Edited K.C. Marshall. Published by Plenum Press, New York, U.S.A..

Palacios C. and F. Gonzalez-Candelas (1997). Analysis of population genetic structure and variability using RAPD markers in the endemic and endangered Limonium dufourii (Plumbaginaceae). Molecular Ecology **6:** 1107-1121.

Pankhurst C.E.,Double B.M., and V.V.S.R Gupta (1998). Biological Indicators of Soil Health.Published by CABI.

Pankhurst C.E., Hawke BG, McDonald HJ, Kirby C.A. Buckerfield J.C. Michelsen P., O'Brien K.A., Gupta V., Doube BM. (1995). Evaluation of soil biological properties as potential bioindicators of soil health. Australian Journal of Experimental Agriculture. **35:** 1015-1028

Priest F. and B. Austin (1993). Modern Bacterial Taxonomy. Chapter 2 Numerical Taxonomy. Published by Chapman and Hall.

Reber H.H. (1992). Simultaneous estimates of the diversity and the degradative capability of heavy-metal-affected soil bacterial communities. Biology and Fertility of Soils **13:** 181-186.

Rosado A.S. , G.F. Duarte, L. Seldin and J.D. Van Elsas (1997). Molecular Microbial Ecology: A mini review. Revista de microbiologia **28:** 135-147.

Rossello-Mora, R. & Amann R. (2001). The species concept for prokaryotes. FEMS Microbiology Reviews, **25:** 39-67.

Rovira (1965). Interaction between plants and roots and soil microorganisms. Archives of Microbiology. **9:** 241-266.

Roberts M, Rothamsted method based on Williams D. E. (1948). Proc. Soil Sci. Soc. Am. **13:** 127.

Saiki R.K., Gelfand D.H., Stoffel S., Scharf S.J., Higuchi R., Horn G., Mullis K. and H.A. Erlich (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487-491.

Sambrook J., Fritsch E.F. and T. Maniatis (1989). Molecular cloning: a laboratory manual, 2nd ed. Cold Harbour Laboratory, Cold Spring Harbour, N.Y.

Schneider M. and F.J. de Bruijn (1996). Rep-PCR mediated genomic fingerprinting of rhizobia and computer-assisted phylogenetic pattern analysis. World Journal of Microbiology and Biotechnology. **12:** 163-174.

Shannon C.E. (1948). A mathematical theory of communication. Bell Syst Technol **27:** 379-423.

Sharp P.A. (1973). Detection of 2 Restriction Endonuclease Activities in Hemophilus-Parainfluenzae Using Analytical Agarose-Ethidium Bromide Electrophoresis Biochemistry, **12:** 3055-3068.

Simpson E.H (1949). Measurement of diversity. Nature **163:** 688.

Sneath P.H.A. and Sokal R.R (1973). Numerical Taxonomy: The principle and practice of numerical classification, W.H. Freeman, San Francisco.

Stackebrandt E. and Goodfellow M. (1991). Nucleic Acid techniques in bacterial systematics. Published by CABS.

Suggs S.V. , T. Hirose, T. Miyake, E.H. Kawashima, M.J. Johnson, K. Itakura, R.B. Wallace (1981). Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences. In: Brown D. D. (Ed.) Developmental Biology Using Purified Genes. New York, Academic Press. 683-693.

Tate R.L. (1995). Soil Microbiology: Chapter 7, The Rhizosphere/ Mycorrhizosphere and Chapter 2, The Soil Ecosystem: Biological Participants. Published by John Wiley and Sons Inc.

Tate R.L. (1995). Soil microbial diversity research: Whither to now? Soil Science **162:** 605-606.

Thomashow L.S, Weller D.M. (1990). Role of Antibiotics and Siderophores in Biocontrol of Take-All Disease of Wheat. Plant Soil **129:** 93-99.

Topal M.D. and Fresco J.R (1976). Complementary base pairing and the origin of substitution mutations. Nature **263:**285-9

Torsvik V. , R. Sorheim and J. Goksoyr (1996). Total bacterial diversity in soil and sediment communities- a review. Journrnal of Industrial Microbiology **17:** 170- 178.

Trevors J.T. (1998). Bacterial biodiversity vin soil with an emphasis on chemically-contaminated soils. Water Air, and Soil Pollution **101:** 45-67.

Turner P. , K.E. Giller, A.M. Chaudri and S. P. Mc Grath.(1995). Population sizes and diversity of Rhizobium are reduced by long-term heavy metal contamination of soil. In:3 rd International Conferences Biogeochemistry of Trace Elements 15-19 May 1995 Abstract no. B1.

Vandamme P. , B. Pot, M. Gillis, P. De Vos, K. Kersters and J. Swings (1996). Polyphasic Taxonomy, a consensus approach to bacterial systematics. Microbiological Reviews **60:** 407-438

Van Elsas, J.D., Trevors, J.T. & Wellington, E.M.H. (1997). *Modern Soil Microbiology.* Marcel Dekker, New York.

Vaneechoulte M., Rossau R., DeVos P., Gillis M., and K. Kersters (1992). Rapid identification of bacteria of the *Comamonadaceae* with amplified ribosomal DNA restriction analysis (ARDRA). FEMS microbial letters. **93:** 227-234

Versalovic J., T. Koeuth and J.R. Lupski (1991). Distribution of repetitive DNA sequences in eubacteria and application of bacterial genomes. Nucleic Acids Research **19:** 6823-6831.

Versalovic J., Schneider, M., de Bruijn, F.J., and Lupski, J.R. (1994). Genomic fingerprinting of bacteria using repetitive sequence based PCR (rep-PCR). Meth. Cell. Mol. Biol. **5:** 25-40

Versalovic J., Kapur V., Mason E.O , Shah U., Koeuth T., Lupski, J.R. (1995). DNA fingerprinting of pathogenic bacteria by fluorophore enhanced repetitive sequence based PCR. Arch. Pathol. Lab. Med. **119:** 23-29

Watters, D.F. & Sweetman, I.C. Rukuhia (1955). Pattern Soil Grinding Machine. Soil Science, **79:** 411-413.

Watkinson A.R. (1998). The role of the soil community in plant population dynamics. Tree **13:** 171-172.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E, Stackebrandt, E., Starr, M.P. & Truper H.G. (1987). Report of the ad-hoc-committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic Bacteriology, **37:** 463.

Winding A. (1994). Beyond the biomass, compositional and functional analysis of soil microbial communities: Chapter 9, Fingerprinting bacterial soil communities using Biolog™ microtitre plates published by John Wiley and Sons Inc.

Weiseburg, W.G., Barns, S.M., Pelletier, D.A. & Lane, D.J. (1991). 16S ribosomal DNA amplification for phylogenetic study. Journal of bacteriology **173:** 687-703.

Wenderoth D.F and Reber H.H., (1999).Development and Comparison of methods to estimate the catabolic versitility of metal-affected soil communities. Soil Biol. Biochem. **31:** 1793-1802

Widmer, F., Seidler, R.J., Gillevet, P.M., Watrud, L.S. & Di Giovanni, G.D. (1998). A highly selective PCR protocol for detecting 16S rRNA genes of the genus *Pseudomonas* (sensu stricto) in environmental samples. Applied and Environmental Microbiology **64:** 2545-2553

Woese, C.R. (1987). Bacterial evolution. Microbiological Reviews, **51:** 221-271.

Young, J.M. (2001). Implications of alternative classifications and horizontal gene transfer for bacterial taxonomy. International Journal of Systematic and Evolutionary Microbiology. **51:** 945-953.

Yin B, Crowley D, Sparovek G, De Melo WJ, and Borneman J (2000). Bacterial Functional Redundancy . Applied and Environmental Microbiology. **66:** 4361-4365

Zak. J, Willig M., Moorhead D., and Wildman H (1994). Functional Diversity of Microbial Communities: A Quantitative Approach. Soil Biol. Biochem. **26:** 1101-1108.

APPENDICES

Appendix A:
Computer Programs

List of Computer Programs

1) Dissimilarity Matrix Program, (1- Jaccard Equation). (see above)

2) Dice Program (see above)

3) Confident Interval Matrix Program (see above)

4) Dendrogram Program (see above)

5) BIOLOG™ 96 : This programs uses the raw data of the 96 well BIOLOG plate and allows for suite of analyses to be done.

6) EcoMicroPlates™ 32: This programs uses the raw data of the 32 well BIOLOG plate and allows for suite of analyses to be done.

7) ERIC Program: A simple matching program for the fingerprint data.

8) Fingerprint Database : A tool for analysing the Fingerprinting data from Genescan.

A1.1: Dissimilarly program (1-jaccard)

```
CALC Num = NLEVELS(Number)
CALC SumB = NLEVELS(SumBase)
VARI [nval=Num] Present[1...SumB]
CALC SB = SumB-1
CALC NM = Num-1
EQUATE [OLDF=!((((1,#SB(-1))#Num),-1)#NM,((1,#SB(-1))#NM),1)] OLD = present;
NEW=Present
GETATT [labels] Number; lab
DELETE [REDEFINE = yes] Gsim
SYMMETRICMATRIX [rows=lab[1]] Gsim
```

```
 FSIMILARITY [SIMILARITY=Gsim] Present[1...SumB]; TEST=jaccard
 " This is to calculate 1 - jaccard "
 DELETE [REDEFINE = yes] G
 SYMMETRICMATRIX [rows=lab[1]] G
 CALC G = 1 - Gsim
 " Creates 3 Symmetrical Matrices: Glower holds the lower value for the Grothues Tummler
   confidence intervals, Gupper holds the upper value of the confidence interval and Ggroups
   indicates which values are grouped/ungrouped with regards to the Grothues Tummler
   confidence interval; 1 - grouped, -1 ungrouped.
 "
 SCALAR TotalA,TotalB; 0
 SYMM [rows=lab[1]] Glower, Gupper, Ggroups
 CALC Ggroups = 0
 SCALAR f[1...SumB]
 VARIATE [nval = Num] old
 for i=1...Num
  for j = i...Num
 " Calculate the Totals for each column "
    CALC old = -1
    CALC old$[i] = 1
    EQUATE [OLDF = old] OLD = Present; f
    CALC TotalA = VSUM(f)
    CALC old$[i] = -1
    CALC old$[j] = 1
    EQUATE [OLDF = old] OLD = Present; f
    CALC TotalB = VSUM(f)

 "  Calculate the confidence interval  "
    CALC AB = (TotalA * TotalB)/SumB
    CALC D = 2*SQRT(( (TotalA*TotalB) * (SumB - (TotalA + TotalB)/2)
)/(SumB*SumB))
    CALC E = 2/(TotalA + TotalB)
    CALC lower = E*(AB - D)
    CALC upper = E*(AB + D)
    CALC Glower$[i;j] = lower
    CALC Gupper$[i;j] = upper
    IF ( (Gsim$[i;j] - upper) .GT. 0)
     CALC Ggroups$[i;j] = 1
    ELSIF ( (Gsim$[i;j] - lower) .LT. 0)
     CALC Ggroups$[i;j] = -1
    ENDIF
  endfor
```

A1.2: Dice Program (1-Dice)

```
DELETE [redefine = yes] DICE
PROCEDURE 'DICE'
PARAMETER          'DATA', 'N','S','MATRIX'; \
                mode = p; \
```

```
                set = 3(yes), no; \
                declared = 3(yes), no; \
                type = 'pointer',2('scalar'), 'symmetricmatrix'; \
                present = 3(yes), no

 VARI [nval=N] p[1...S]
 CALC SB = S-1
 CALC NM = N-1
 matr [N;S] xt
 equate  [OLDF=!( ((1,#NM(-1))#S,-1)#NM,((1,#NM(-1))#SB),1)] DATA; xt
 pointer [N] y
 matr [N;S] y[]
 for i=1...N
  calc a = (i - 1)*S
  calc b = (N - i)*S
  equate [OLDF=!(  #a(-1),#S(1),#b(-1)  )] xt;y[i]
 endfor
 matrix [S;1] v
 calc v = 1
 matrix [N;N] tempmat
 calc tempmat$[*;1...N] =  2*( ((y[].eq.xt).AND.(y[].ne.0)) *+ v ) / \
      (2*( ((y[].eq.xt).AND.(y[].ne.0)) *+ v) + \
      ( (y[].ne.xt) *+ v) )
 symm [N] MATRIX
 calc MATRIX = tempmat
 ENDPROCEDURE
```

A1.3 Cofident Interval Matrix Program

```
CALC Num = NLEVELS(Number)
CALC SumB = NLEVELS(SumBase)
VARI [nval=Num] Present[1...SumB]
CALC SB = SumB-1
CALC NM = Num-1
EQUATE [OLDF=!((((1,#SB(-1))#Num),-1)#NM,((1,#SB(-1))#NM),1)] OLD = present;
NEW=Present
GETATT [labels] Number; lab
DELETE [REDEFINE = yes] Gsim
SYMMETRICMATRIX [rows=lab[1]] Gsim
FSIMILARITY [SIMILARITY=Gsim] Present[1...SumB]; TEST=jaccard

" This is to calculate 1 - jaccard "
DELETE [REDEFINE = yes] G
SYMMETRICMATRIX [rows=lab[1]] G
CALC G = 1 - Gsim

" Creates 3 Symmetrical Matrices: Glower holds the lower value for the Grothues Tummler
 confidence intervals, Gupper holds the upper value of the confidence interval and Ggroups
 indicates which values are grouped/ungrouped with regards to the Grothues Tummler
```

```
  confidence interval; 1 - grouped, -1 ungrouped.
 "
 SCALAR TotalA,TotalB; 0
 SYMM [rows=lab[1]] Glower, Gupper, Ggroups
 CALC Ggroups = 0
 SCALAR f[1...SumB]
 VARIATE [nval = Num] old
 for i=1...Num
  for j = i...Num
 " Calculate the Totals for each column "
    CALC old = -1
    CALC old$[i] = 1
    EQUATE [OLDF = old] OLD = Present; f
    CALC TotalA = VSUM(f)
    CALC old$[i] = -1
    CALC old$[j] = 1
    EQUATE [OLDF = old] OLD = Present; f
    CALC TotalB = VSUM(f)

 "  Calculate the confidence interval  "
    CALC AB = (TotalA * TotalB)/SumB
    CALC D = 2*SQRT(( (TotalA*TotalB) * (SumB - (TotalA + TotalB)/2)
)/(SumB*SumB))
    CALC E = 2/(TotalA + TotalB)
    CALC lower = E*(AB - D)
    CALC upper = E*(AB + D)
    CALC Glower$[i;j] = lower
    CALC Gupper$[i;j] = upper
    IF ( (Gsim$[i;j] - upper) .GT. 0)
     CALC Ggroups$[i;j] = 1
    ELSIF ( (Gsim$[i;j] - lower) .LT. 0)
     CALC Ggroups$[i;j] = -1
    ENDIF
  endfor
```

A1.4: Dendrogram Program

```
delete [redefine=y]
ddendrogram,DDEN_RENODE,DDEN_ORDER,DDEN_BUILD,DDEN_DISPLAY,DDEN
_GRAPH
 " Procedures for drawing high-resolution dendrograms "
 "DDENDROGRAM calls
DDEN_RENODE,DDEN_ORDER,DDEN_BUILD,DDEN_DISPLAY,DDEN_GRAPH"
 " - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
 Procedure 'DDEN_RENODE'
 " Pete Digby, Rothamsted Experimental Station, 24.6.88
  Examines hcluster amalgamations data: renodes it, stores 'proper' node
  numbers, similarities, group sizes, and ziggurat-degree in EXTRA "
```

```
parameter 'AMALGAM',  " (I: matrix) amalgamations data " \
        'EXTRA';    " (O: matrix) extra information " \
        mode = p,p
calc     n2m1 = 2 * (n = (nm1 = nrow(AMALGAM)) + 1) - 1
variate  [nm1] m[1,2],t[1,2],gs,s[1,2],sel; (*)5,(!((1)#nm1))3
 &       [n2m1; val=(0)#n2m1] zd
scalar   i,lm,rm,lz,rz,node,wn,zm,zeq; 0,(*)8
calc     m[1,2] = AMALGAM$[*;1,2]
  " loop through nodes: set new node numbers, sizes, zig. degree "
for      [ntimes=nm1]
 calc     node = n + (i = i + 1)
  &       sel$[i] = 0
  &       lm,rm = (m[])$[i]
  &       wn = s[1]$[i] + s[2]$[i]
  &       t[] = sel * (m[] == lm)
  &       m[],s[] = m[],s[] * (1 - t[]) + 2(node,wn) * t[]
  &       t[] = sel * (m[] == rm)
  &       m[],s[] = m[],s[] * (1 - t[]) + 2(node,wn) * t[]
  &       gs$[i] = wn
  &       lz,rz = zd$[lm,rm]
  &       zm = lz * (lz >= rz) + rz * (rz > lz)
  &       zeq = (int(lz / n) == int(rz / n))
  &       zd$[node] = (zm + 1) * (1 - zeq) + zeq * n * (int(zm / n) + 1)
endfor
matrix   [rows=nm1; columns= \
        !T(L_node,R_node,Similrty,L_size,R_size,G_size,Z_degree)] EXTRA
calc     EXTRA$[*; 1,2] = m[]
 &       EXTRA$[*; 3] = AMALGAM$[*; 3]
 &       EXTRA$[*; 4...6] = s[],gs
 &       sel = n + cum(sel = 1)
 &       EXTRA$[*; 7] = integer(zd$[sel] / n)
endproc
" - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
procedure 'DDEN_ORDER'
" Pete Digby, Rothamsted Experimental Station, 24.6.88
 Returns ordering for hierarchical cluster analysis "
option   'METHOD', " (I: string {ZIGGURAT, SIZE, FIRST, SWITCH} default
                ZIGGURAT) (primary) ordering method to use " \
        'SECOND'; " (I: string {ZIGGURAT, SIZE, FIRST} default SIZE)
                secondary method to resolve indeterminacies " \
        mode = t,t; \
        default = 'ZIGGURAT','FIRST'
parameter 'EXTRA',  " (I: matrix) extra information, from DDEN_REN " \
        'ORDER',  " (O: variate) output ordering " \
        'SWITCH'; " (I: variate) supplied switching, if required " \
        mode = p,p,p
scalar   n,nm1,n2m1,pmn,pmnp1,lm,rm,thism
calc     n2m1 = 2 * (n = (nm1 = nrows(EXTRA)) + 1) - 1
variate  [n] ORDER,tord,nv; !((0)#n),*,!(1...n)
variate  [nm1] sel,nm1v,m[1,2],s[1,2],z[1,2],swz,swf,sws,sw
```

```
    &       [n2m1; val=(0)#n2m1] zd
  calc      sel = (nm1v = cum(nm1v = 1)) + n
    &       zd$[sel] = EXTRA$[*; 7]
    &       m[],s[] = EXTRA$[*; 1,2,4,5]
    &       z[] = zd$[m[]]
     " get switch variable "
  if        METHOD .eqs. 'SWITCH'
    if        unset(SWITCH)
      calc      sw = 0
    else
      calc      sw = SWITCH
    endif
  else
     " form switch variable, depending on options "
    calc      swz,swf,sws = (z[1],m[2],s[1] < z[2],m[1],s[2]) - 0.5  \
                    + (z[1],m[2],s[1] == z[2],m[1],s[2]) / 2
    if        METHOD .eqs. 'ZIGGURAT'
      if        SECOND .eqs. 'FIRST'
        calc      sw = 2 * swz + swf
      else
        calc      sw = 4 * swz + 2 * sws + swf
      endif
    elsif     METHOD .eqs. 'FIRST'
      calc      sw = swf
    elsif     SECOND .eqs. 'ZIGGURAT'
      calc      sw = 4 * sws + 2 * swz + swf
    else
      calc      sw = 2 * sws + swf
    endif
    calc      sw = ((sw / abs(sw)) + 1) / 2
  endif
    " switch m[1,2] as required, then initialise order "
  calc      mt = vsum(m)
    &       m[2] = mt - (m[1] = m[2] * sw + m[1] * (1 - sw))
  calc      ORDER$[1] = n2m1
    " loop expanding nodes "
  for       [ntimes=nm1]
    calc      pmnp1 = (pmn = max(nv * (ORDER == (thism = max(ORDER))))) + 1
      &       thism = thism - n
      &       sel = 1
      &       sel$[pmn] = 2
      &       sel = cum(sel)
      &       tord$[sel] = ORDER$[nm1v]
      &       lm,rm = m[1]$[thism],m[2]$[thism]
      &       tord$[pmn,pmnp1] = lm,rm
      &       ORDER = tord
  endfor
  endproc
  " - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
  Procedure 'DDEN_BUILD'
```

```
" Pete Digby, Rothamsted Experimental Station, 24.6.88
  Prepares dendrogram information for DDENDRO "
option   'STYLE',      " (I: string {AVERAGE, CENTROID, FULL, LOWER}
                       default AVERAGE) style of drawing for links " \
         'IN';         " (I: string {SIMILARITES, PERCENTAGES, DISTANCES}
                       default SIMILARITIES} form of input " \
         mode = t,t; \
         default = 'AVERAGE','SIMILARITIES'
parameter 'EXTRA',     " (I: matrix) extra amalgamation data " \
         'PERMUTATION', " (I: variate) permutation of units " \
         'DENDRO';     " (O: pointer) dendrogram information " \
         mode = p,p,p
scalar   n,nm1,n2m1,old[1,2],oldx[1,2],oldy[1,2],  \
         newx,newy,hy[1,2],ly[1,2],size[1,2],i,npi
calc     n2m1 = 2 * (n = (nm1 = nrows(EXTRA)) + 1) - 1
pointer  [nm1] posx,posy
variate  [4] posx[],posy[]
pointer  [val=posx,posy] DENDRO
variate  [n2m1] nodex,nodey,hiy,loy
calc     nodey$[PERMUTATION] = (!(n...1) - 0.5) / n
 &       hiy = (loy = nodey)
 &       nodex$[PERMUTATION] = 0
 &       i = 0
for      [ntimes=nm1]
 calc      npi = n + (i = i + 1)
  &        old[] = EXTRA$[i; 1,2]
  &        oldx[] = nodex$[old[]]
 if        IN .eqs. 'SIMILARITIES'
  calc       newx = 1 - EXTRA$[i; 3]
 elsif     IN .eqs. 'PERCENTAGES'
  calc       newx = 100 - EXTRA$[i; 3]
 else
  calc       newx = EXTRA$[i; 3]
 endif
 calc      nodex$[npi] = newx
  &        posx[i] = !(oldx[1],newx,newx,oldx[2])
 if        STYLE .eqs. 'FULL'
  calc       hy[] = hiy$[old[]]
   &         ly[] = loy$[old[]]
  if         hy[1] > hy[2]
   calc        hiy$[npi] = (oldy[1] = hy[1])
    &          loy$[npi] = (oldy[2] = ly[2])
  else
   calc        loy$[npi] = (oldy[1] = ly[1])
    &          hiy$[npi] = (oldy[2] = hy[2])
  endif
 else
  calc       oldy[] = nodey$[old[]]
  if         STYLE .eqs. 'CENTROID'
   calc        size[] = EXTRA$[i; 4,5]
```

```
      &      vsum = vsum(size)
      &      newy = (oldy[1] * size[1] + oldy[2] * size[2]) / vsum
   elsif    STYLE .eqs. 'LOWER'
    calc     newy = vmin(oldy)
   else
    calc     newy = vmean(oldy)
   endif
   calc     nodey$[npi] = newy
 endif
 calc     posy[i] = !(2(oldy[]))
endfor
endproc
" - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
PROCEDURE 'DDEN_DISPLAY'
" Draws dendrograms using high-resolution graphics "
option   'TITLE',    " title for graph " \
         'WINDOW',   " window for graph " \
         'SCREEN',   " setting for screen option of dgraph " \
         'METHOD',   " form of input (e.g. distances) " \
         'SETSCALE', " whether to set scale for similarity axis " \
         'ORIENT',   " orientation of dendrogram " \
         'REVERSE';  " whether to reverse direction of units " \
         mode = t,p,t,t,t,t,t; \
         default = ' ',1,'CLEAR','SIMILARITIES','NO','WEST','NO'
parameter 'DENDRO',   " information on dendrogram (from DDEN_GET) " \
         'PERMUTATION', " order of units " \
         'LABELS',   " labels for units " \
         'PENS';     " pens to use for links " \
         mode = p
scalar   nm1,n,xyeqxy,maxx,yl,yh,xl,xh,yo,xo
scalar   typvar; 4
calc     n = (nm1 = nval(DENDRO[1])) + 1
pointer  [nm1] gx,gy
variate  [4] gx[],gy[]
  " get max. x "
if       (METHOD .eqs. 'PERCENTAGES') .and. (SETSCALE .eqs. 'YES')
 calc     maxx = 100
elsif    (METHOD .eqs. 'SIMILARITIES') .and. (SETSCALE .eqs. 'YES')
 calc     maxx = 1
else
 calc     maxx = max(vmax(DENDRO[1]))
endif
  " get x,y coords. to plot, and axis ranges "
if       xyeqxy = (ORIENT .eqs. 'WEST') .or. (ORIENT .eqs. 'EAST')
 calc     gx[],gy[] = #DENDRO
  &       yl,yh,yo = 0,1
 if       ORIENT .eqs. 'EAST'
  calc     gx[] = maxx - gx[]
   &       xl = 0
   &       xo = (xh = maxx * (1 + 0.20))
```

```
 else
  calc     xh = maxx
   &       xo = (xl = -0.05 * maxx)
 endif
 if       REVERSE .eqs. 'YES'
  calc     gy[] = 1 - gy[]
 endif
else
 calc     gy[],gx[] = #DENDRO
  &       xl,xh,xo = 0,1
 if       ORIENT .eqs. 'NORTH'
  calc     gy[] = maxx - gy[]
   &       yl = 0
   &       yo = (yh = maxx * (1 + 0.15))
 else
  calc     yh = maxx
   &       yo = (yl = -0.05 * maxx)
 endif
 if       REVERSE .eqs. 'YES'
  calc     gx[] = 1 - gx[]
 endif
endif
  " sort out axis ranges and unit-axis labels "
axes WINDOW; ylo=yl; yh; xl; xh; yor=yo; xo
if       unset(LABELS)
 ASSIGN !(1...n); LABELS
endif
getattrib [type] LABELS; plabs
if       plabs['type'] == typvar
 print    [ch=ulabs; rlw=0; squash=y; iprint=*] LABELS; 1; 0
else
 text     ulabs; LABELS
endif
if       REVERSE .eqs. 'NO'
 calc     ulat = (!(n...1) - 0.5) / n
else
 calc     ulat = (!(1...n) - 0.5) / n
endif
sort     [PERMUTATION] ulat
  " Fix to draw the Levels; Added by DAM 20/1/98 "
if    (METHOD .eqs. 'SIMILARITIES')
 if       xyeqxy
  axes     WINDOW; ymark=ulat; ylab=ulabs; xtitle='Levels'; \
               xmark=!(0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0); \
               xlab=!t('100','90','80','70','60','50','40','30','20','10','0'); style=xy
 else
  axes     WINDOW; xmark=ulat; xlab=ulabs; ytitle='Levels'; \
               ymark=!(0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0); \
               ylab=!t('100','90','80','70','60','50','40','30','20','10','0'); style=xy
 endif
```

```
 else
" Original Code"
  if      xyeqxy
   axes     WINDOW; ymark=ulat; ylab=ulabs; style=y
  else
   axes     WINDOW; xmark=ulat; xlab=ulabs; style=x
  endif
 endif

   " sort out pens to use "
 if      unset(PENS)
  assign   1; PENS
 elsif    nvalues(PENS) > 1
  groups   PENS; factor=temp; levels=penlist
  "sort    [PENS; groups=temp; levels=penlist]"
  pen      #penlist; method=line; symbol=0; join=given
 endif
 if      nvalues(PENS) == 1
  pen      PENS; method=line; line=1; symbol=0; join=given
 endif
 dgraph [TITLE=#TITLE; WINDOW=WINDOW; KEYWINDOW=0;
SCREEN=#SCREEN] \
      gy[]; gx[]; #PENS
 endproc
" - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
 PROCEDURE 'DDEN_GRAPH'
"  Draws dendrograms using line-printer graphics  "
 option   'TITLE',    " title for graph " \
         'NCOLUMNS', " no. columns for graph " \
         'METHOD',   " form of input (e.g. distances) " \
         'SETSCALE', " whether to set scale for similarity axis " \
         'REVERSE';  " whether to reverse direction of units " \
         mode = t,p,t,t,t; \
         default = ' ',61,'SIMILARITIES','NO','NO'
parameter 'DENDRO',   " information on dendrogram (from DDEN_GET) " \
         'PERMUTATION', " order of units " \
         'LABELS',   " labels for units " \
         'SYMBOLS';  " symbols to use for links " \
         mode = p
scalar   nm1,n,maxx
scalar   typvar; 4
calc     n = (nm1 = nval(DENDRO[1])) + 1
pointer  [nm1] gx,gy
variate  [4] gx[],gy[]
  " get max. x "
if      (METHOD .eqs. 'PERCENTAGES') .and. (SETSCALE .eqs. 'YES')
 calc     maxx = 100
elsif    (METHOD .eqs. 'SIMILARITIES') .and. (SETSCALE .eqs. 'YES')
 calc     maxx = 1
else
```

```
 calc     maxx = max(vmax(DENDRO[1]))
endif
  " get x,y coords. to plot, and axis ranges "
calc     gx[],gy[] = #DENDRO
if       REVERSE .eqs. 'YES'
 calc     gy[] = 1 - gy[]
endif
calc     nrows = 2 * n + 1
  " sort out unit-axis labels "
if       unset(LABELS)
 ASSIGN !(1...n); LABELS
endif
getattrib [type,nvalues] LABELS; plabs
if       plabs['type'] == typvar
 print    [ch=ulabs; rlw=0; squash=y; iprint=*] LABELS; 1; 0
elsif    plabs['nvalues'] == n
 text     ulabs; LABELS
else
 text     ulabs; !T((' ')#n)
endif
if       REVERSE .eqs. 'NO'
 calc     ulat = (!(n...1) - 0.5) / n
else
 calc     ulat = (!(1...n) - 0.5) / n
endif
sort     [PERMUTATION] ulat
  " sort out symbols to use "
if       unset(SYMBOLS)
 text     gsymbols; !T(('+')#nm1)
elsif    nvalues(SYMBOLS) == 1
 text     gsymbols; !T((#SYMBOLS)#nm1)
else
 text     gsymbols; !T(#SYMBOLS)
endif
graph [TITLE=#TITLE; YLOWER=0; 1; -0.2; maxx; JOIN=given; \
 NROWS=nrows; NCOLUMNS; YINTEGER=yes] \
 gy[],#ulat; gx[],(-0.18)#n; (l)#nm1,(t)#n; #gsymbols,#ulabs
endproc
" - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - "
Procedure [RESTORE=dsave] 'DDENDROGRAM'
" Pete Digby, Rothamsted Experimental Station, 24.6.88
 Draws dendrograms using high-resolution graphics "
OPTION NAME= \
 'STYLE',      " (I: string {average, centroid, full, lower}; default
               average) Style to use for the links of the dendrogram " \
 'ORDERING',   " (I: strings {given, ziggurat, size, first} default
               ziggurat,size,first) How to define the order of the units
               for the dendrogram " \
 'REVERSE',    " (I: string {no, yes} default no) Whether to reverse the
               order of the units in the dendrogram " \
```

```
 'ORIENTATION', " (I: string {west, north, south, east} default west)
          Specifies the orientation of the dendrogram " \
 'SETSCALE',   " (I: string {no, yes} default no) Whether the procedure
          should set the scale for the axis showing similarity, or
          whether the scale should be determined by the range of
          similarities " \
 'METHOD',     " (I: string {similarities, percentages, distances} default
          similarities) Method used to represent the scale on which
          the amalgamations have been made " \
 'SCREEN',     " (I: string {clear, keep} default clear) Setting to use
          for the SCREEN option of DGRAPH " \
 'CHANGE',     " (I: string {order, dendrogram, display} default order)
          If a dendrogram-save structure from a previous DDENDROGRAM
          is used as the DATA parameter then this option specifies the
          area of the process where the first changes occur " \
 'GRAPHICS';   " (I: string {highresolution, lineprinter} default line)
          Which type of graphics to use " \
 MODE=t,t,t,t,t,t,t,t,t; \
 DEFAULT='average',!T(ziggurat,size,first),'no','west', \
      'no','similarities','clear','order','highresolution'; \
 LIST=no,yes,7(no)
PARAMETER NAME= \
 'DATA',       " (I: matrix or pointer) Data to form each dendrogram " \
 'PERMUTATION', " (I/O: variate) Specify or save order of the units " \
 'LABELS',     " (I: text or variate) Labels to use for the units " \
 'TITLE',      " (I: text) Title for the dendrogram " \
 'WINDOW',     " (I: scalar) Window to use; or zero to suppress dendrogram;
          for lineprinter graphics windows 1,2 gives 101,61 columns
          for graph (window 1 if unset) " \
 'PENS',       " (I: scalar or variate) Graphical pen(s) to use (pen 1 if
          unset), LINESTYLE=1 set if not variate " \
 'ZIGGURAT',   " (O: variate) Save the 'ziggurat-degree' of the links " \
 'SAVE';       " (O: pointer) Save the dendrogram information " \
 MODE=p,p,p,p,p,p,p,p

SCALAR Tvariate,Tmatrix,Tpointer; VALUE=4,5,14
"  basic checks on input structure, get (most) input string options  "
EXIT [CONTROL=proc; EXPLANATION='*****  DATA parameter must be supplied'] \
    UNSET(DATA)
&   [EXPLANATION=!t('***** values of DATA parameter all missing',\
    '(amalgamations cannot be produced from single linkage clustering;',\
    'the minimum spanning tree, from HDISPLAY, must be used instead)')] \
    NMV(DATA)==NVALUES(DATA)
GETATTRIBUTE [ATTRIBUTE=type,nrows,ncolumns,nmv] DATA; Pd
CALC DSaveIn = Pd['type'] == Tpointer
IF (.not. DSaveIn) .and. (Pd['type'] <> Tmatrix)
 PRINT \
'*****  DATA parameter must be a matrix, or a DDENDROGRAM save structure'
 EXIT [CONTROL=procedure]
ENDIF
```

```
CALC OScrKeep,OSScYes,OOrEast,OOrNorth,OOrSouth,ORevYes, \
    OStyFull,OStyLow,OStyCent,OInPerc,OInDist,OChgDen,OChgDis,OGrLP = \
    sum(SCREEN,SETSCALE,(ORIENTATION)3,REVERSE, \
      (STYLE)3,2(METHOD,CHANGE),GRAPHICS .in. \
  !T(KEEP,keep,K,k,KE,ke,KEE,kee), \
  !T(YES,yes,Y,y,YE,ye), \
  !T(EAST,east,E,e,EA,ea,EAS,eas), \
  !T(NORTH,north,N,n,NO,no,NOR,nor,NORT,nort), \
  !T(SOUTH,south,S,s,SO,so,SOU,sou,SOUT,sout), \
  !T(YES,yes,Y,y,YE,ye), \
  !T(FULL,full,F,f,FU,fu,FUL,ful), \
  !T(LOWER,lower,L,l,LO,lo,LOW,low,LOWE,lowe), \
  !T(CENTROID,centroid,C,c,CE,ce,CEN,cen,CENT,cent,CENTR,centr, \
    CENTRO,centro,CENTROI,centroi), \
  !T(PERCENTAGES,percentages,P,p,PE,pe,PER,per,PERC,perc,PERCE,perce, \
    PERCEN,percen,PERCENT,percent,PERCENTA,percenta,PERCENTAGE,percentage),
\
  !T(DISTANCES,distances,D,d,DI,di,DIS,dis,DIST,dist,DISTA,dista, \
    DISTAN,distan,DISTANC,distanc,DISTANCE,distance), \
  !T(DENDROGRAM,dendrogram,D,d,DE,de,DEN,den,DEND,dend,DENDR,dendr, \
    DENDRO,dendro,DENDROG,dendrog,DENDROGR,dendrogr), \
  !T(DISPLAY,display,DI,di,DIS,dis,DISP,disp,DISPL,displ,DISPLA,displa), \
  !T(LINEPRINTER,lineprinter,L,l,LI,li,LIN,lin,LINE,line,LINEP,linep, \
    LINEPR,linepr,LINEPRI,linepri,LINEPRIN,lineprin)) > 0
"  Preliminary - get from input matrix to EXTRA from RENODE,
            or ASSIGN EXTRA from input DDENDROGRAM Save structure  "
IF .not. DSaveIn
"  input is matrix - preliminary processing: MST to AMAL, renode  "
  IF Pd['ncolumns'] == 2
    CALC     Num1 = (Nu = Pd['nrows']) - 1
    VARIATE  [Num1] Amal_l,Amal_r,Amal_s; !(2...Nu),(*)2
    CALC     Amal_r,Amal_s = DATA$[Amal_l; 1,2]
    SORT     [DIRECTION=descending] Amal_s,Amal_l,Amal_r
    MATRIX   [Num1;3] AMALGAM
    CALC     AMALGAM$[*;1...3] = Amal_l,Amal_r,Amal_s
  ELSIF Pd['ncolumns'] == 3
    CALC     Nu = (Num1 = Pd['nrows']) + 1
    DUMMY    AMALGAM
    ASSIGN   DATA; AMALGAM
  ELSE
    PRINT '*****  DATA matrix does not have 2 or 3 columns'
    EXIT [CONTROL=procedure]
  ENDIF
  " renode amalgamations, get extra information "
  DDEN_RENODE AMALGAM; EXTRA
ELSE
"  input is DDENDROGRAM Save structure  "
  DUMMY EXTRA
  ASSIGN DATA[1]; EXTRA
  CALC Nu = (Num1 = nrows(EXTRA)) + 1
```

```
ENDIF
"  Ordering stage - required from input Save if CHANGE='order',
                or from input matrix  "
IF .not. (DSaveIn .and. (OChgDen .or. OChgDis))
"  sort out ordering option and permutation parameter  "
 "  check for 'given' option setting "
 IF OOrdGive = sum(ORDERING .in. \
     !T(GIVEN,given,G,g,GI,gi,GIV,giv,GIVE,give)) > 0
  "  check for OK parameter "
  IF unset(PERMUTATION)
   PRINT '*****  PERMUTATION parameter not supplied for ORDERING=given'
   EXIT [CONTROL=procedure]
  ELSE
   GETATTRIBUTE [ATTRIBUTE=type,nvalues,nmv] PERMUTATION; Pperm
   IF (Pperm['type'] <> Tvariate) .or. (Pperm['nvalues'] <> Nu)
    PRINT '*****  PERMUTATION parameter not variate of correct length'
    EXIT [CONTROL=procedure]
   ELSIF sum(sort(PERMUTATION) <> !(1...Nu))
    PRINT '*****  PERMUTATION parameter invalid'
    EXIT [CONTROL=procedure]
   ENDIF
  ENDIF
 ELSIF unset(PERMUTATION)
  VARIATE [Nu] DdenPerm
  ASSIGN DdenPerm; PERMUTATION
 ELSE
  GETATTRIBUTE [ATTRIBUTE=type] PERMUTATION; Pperm
  IF (Pperm['type'] <> Tvariate) .and. (nmv(Pperm['type']) == 0)
   PRINT '*****  PERMUTATION save-structure not of type variate'
   EXIT [CONTROL=procedure]
  ELSE
   VARIATE [Nu] PERMUTATION
  ENDIF
 ENDIF
"  do the ordering, if required  "
IF .not. OOrdGive
 "  sort out the option setting(s); only first 2 needed "
 IF .not. unset(ORDERING)
  CALC NOrdSet = 1 + (nvalues(ORDERING) > 1)
  POINTER [NOrdSet] OOrdBits,OOrdZig,OOrdSize,OOrdFrst
  TEXT [1] OOrdBits[]
  EQUATE ORDERING; OOrdBits
  CALC (OOrdZig,OOrdSize,OOrdFrst)[] = \
       sum(3(OOrdBits[]) .in. \
       !T(ZIGGURAT,ziggurat,Z,z,ZI,zi,ZIG,zig,ZIGG,zigg,ZIGGU,ziggu, \
         ZIGGUR,ziggur,ZIGGURA,ziggura), \
       !T(SIZE,size,S,s,SI,si,SIZ,siz), \
       !T(FIRST,first,F,f,FI,fi,FIR,fir,FIRS,firs)) > 0
  IF OOrdSize[1]
   TEXT ORDER1; 'SIZE'
```

```
    ELSIF OOrdFrst[1]
     TEXT ORDER1; 'FIRST'
    ELSE
     TEXT ORDER1; 'ZIGGURAT'
     CALC OOrdZig[1] = 1
    ENDIF
    IF NOrdSet > 1
     IF OOrdSize[2] .and. .not. OOrdSize[1]
      TEXT ORDER2; 'SIZE'
     ELSIF OOrdZig[2] .and. .not. OOrdZig[1]
      TEXT ORDER2; 'ZIGGURAT'
     ELSIF OOrdFrst[2] .and. .not. OOrdFrst[1]
      TEXT ORDER2; 'FIRST'
     ELSIF .not. OOrdSize[1]
      TEXT ORDER2; 'SIZE'
     ELSIF .not. OOrdZig[1]
      TEXT ORDER2; 'ZIGGURAT'
     ELSE
      TEXT ORDER2; 'FIRST'
     ENDIF
    ELSE
     TEXT ORDER2; ' '
    ENDIF
   ELSE
    " provide default ordering criteria "
    TEXT ORDER1,ORDER2; 'ZIGGURAT','SIZE'
   ENDIF
   " do the ordering "
   DDEN_ORDER [#ORDER1; #ORDER2] EXTRA; PERMUTATION
  ENDIF
ELSE
" Ordering from DDENDROGRAM Save structure "
 ASSIGN DATA[2]; PERMUTATION
ENDIF
" check for missing setting of WINDOW "
IF unset(WINDOW)
 ASSIGN 1; WINDOW
ENDIF
" get METHOD option (needed in two places) "
 IF OInPerc
  TEXT OInput; 'PERCENTAGES'
 ELSIF OInDist
  TEXT OInput; 'DISTANCES'
 ELSE
  TEXT OInput; 'SIMILARITIES'
 ENDIF
" Dendrogram stage - required from input Save if CHANGE='dendrogram',
            or from input matrix; provided WINDOW <> 0 "
IF (.not. (DSaveIn .and. OChgDis)) .and. (WINDOW > 0)
" get options and call DDEN_BUILD "
```

```
 IF OStyFull
  TEXT OStyle; 'FULL'
 ELSIF OStyLow
  TEXT OStyle; 'LOWER'
 ELSIF OStyCent
  TEXT OStyle; 'CENTROID'
 ELSE
  TEXT OStyle; 'AVERAGE'
 ENDIF
 DDEN_BUILD [#OStyle; #OInput] EXTRA; PERMUTATION; DENDRO
ELSIF WINDOW > 0
" Link information from DDENDROGRAM Save structure "
 DUMMY DENDRO
 ASSIGN DATA[3]; DENDRO
ELSE
 POINTER [2] DENDRO
ENDIF
" Display stage - required unless WINDOW = 0 "
IF WINDOW > 0
 IF OScrKeep
  TEXT OScreen; 'KEEP'
 ELSE
  TEXT OScreen; 'CLEAR'
 ENDIF
 IF OSScYes
  TEXT OSetScale; 'YES'
 ELSE
  TEXT OSetScale; 'NO'
 ENDIF
 IF OOrEast
  TEXT OOrient; 'EAST'
 ELSIF OOrNorth
  TEXT OOrient; 'NORTH'
 ELSIF OOrSouth
  TEXT OOrient; 'SOUTH'
 ELSE
  TEXT OOrient; 'WEST'
 ENDIF
 IF ORevYes
  TEXT OReverse; 'YES'
 ELSE
  TEXT OReverse; 'NO'
 ENDIF
 IF unset(TITLE)
  ASSIGN ' '; TITLE
 ENDIF
 IF OGrLP
  CALC Ncols = 141 - WINDOW * 40
  DDEN_GRAPH [#TITLE; Ncols; #OInput; #OSetScale; #OReverse] \
   DENDRO; PERMUTATION; LABELS; PENS
```

```
  ELSE
    DDEN_DISPLAY [#TITLE; WINDOW; #OScreen; #OInput; \
          #OSetScale; #OOrient; #OReverse] \
      DENDRO; PERMUTATION; LABELS; PENS
  ENDIF
ENDIF
"  ZIGGURAT and DDENDROGRAM save structures  "
IF .not. unset(ZIGGURAT)
  VARIATE [Num1] ZIGGURAT
  CALC ZIGGURAT = EXTRA$[*; 7]
ENDIF
IF .not. unset(SAVE)
  VARIATE PERMKEEP; PERMUTATION
  POINTER [VALUES=EXTRA,PERMKEEP,DENDRO] SAVE
ENDIF
ENDPROCEDURE
```

All programs can be obtained from my supervisor, Dr Penny Hirsch at Rothamsted Research, Rothamsted Experimental Station, Harpenden HERTS, AL5 5AS email address: penny.hirsch@bbsrc.ac.uk

Appendix B:
BIOLOG™ Data

Principle Component Analysis Scores and the latent Vectors (Loadings) for the BIOLOG GN1experiments.

1) Mar 98 BIOLOG™

PCA Scores					
Plot	First Component	Second Component	Third Component	Fourth Component	Fifth Component
FYM (rep 1)	-1.1401	1.2724	0.2836	-0.0346	-0.6893
FYM (rep 2)	-1.3302	-1.3042	0.6760	-0.1654	0.3232
FYM (rep 3)	0.0757	0.7830	-0.6419	-0.7712	0.7150
Sewage (rep 1)	0.0367	0.1599	-0.4074	1.2659	0.3385
Sewage (rep 2)	1.7943	0.1334	1.0038	-0.0662	-0.0284
Sewage (rep 3)	0.5636	-1.0445	-0.9140	-0.2285	-0.6591

Latent Vectors (Loadings)

	1	2	3	4	5
vsource[2]	-0.01294	-0.01400	-0.01781	0.03254	0.01948
vsource[3]	0.00372	0.08785	-0.09906	-0.17200	0.20587
vsource[4]	0.04404	-0.03100	0.11498	-0.03151	-0.00963
vsource[5]	0.03550	0.00951	-0.03425	-0.02441	-0.04001
vsource[6]	0.01259	-0.00534	-0.02829	-0.03644	0.00787
vsource[7]	-0.00423	0.00172	-0.01310	0.00611	-0.04390
vsource[8]	-0.02121	0.06741	0.06516	-0.04369	-0.10333
vsource[9]	-0.02771	-0.04245	0.02806	-0.01330	0.02925
vsource[10]	0.07610	0.08452	-0.03514	-0.03919	0.01671
vsource[11]	0.00725	0.01251	-0.11015	0.00528	-0.00836
vsource[12]	0.01622	0.00833	-0.01005	0.00126	-0.00429
vsource[13]	-0.00032	0.01091	-0.01390	-0.01877	0.01783
vsource[14]	0.06278	0.03416	-0.06709	-0.11544	-0.01760
vsource[15]	0.02955	-0.00072	-0.03049	-0.01283	0.00005
vsource[16]	-0.07226	-0.11238	0.09775	0.05418	0.04288
vsource[17]	0.02353	-0.00537	-0.01842	-0.02223	-0.02966
vsource[18]	0.27424	0.03892	-0.23796	-0.03969	0.14184
vsource[19]	-0.01613	-0.03959	-0.03344	-0.06419	0.17447

vsource[20]	0.01241	-0.00789	-0.02523	0.01150	-0.01340
vsource[21]	0.00894	-0.00700	-0.01663	-0.00740	-0.02644
vsource[22]	-0.06012	-0.10326	0.09808	-0.02664	0.04892
vsource[23]	0.01571	-0.10198	0.00979	0.03182	-0.07158
vsource[24]	0.02329	-0.01043	0.01168	-0.06442	-0.04491
vsource[25]	0.00967	-0.00567	-0.00897	0.00101	-0.00505
vsource[26]	-0.04038	-0.05200	0.03992	-0.02650	0.03280
vsource[27]	0.00537	0.00753	-0.16969	-0.07116	-0.02300
vsource[28]	0.01508	-0.00436	0.01157	-0.00178	-0.01002
vsource[29]	0.01228	0.00211	-0.01696	-0.01182	-0.02086
vsource[30]	0.08351	0.02764	0.12285	-0.00814	-0.05648
vsource[31]	0.01363	0.10317	-0.02642	-0.12163	-0.07912
vsource[32]	0.06759	0.06898	-0.03184	0.06832	-0.30517
vsource[33]	-0.02818	-0.02694	-0.01242	-0.04996	0.00055
vsource[34]	0.00422	-0.00410	-0.01483	-0.00104	-0.02181
vsource[35]	0.00298	-0.00580	-0.03094	0.04528	0.00659
vsource[36]	0.00617	-0.00098	-0.01085	-0.01398	0.01686
vsource[37]	0.01029	0.00175	-0.01467	0.00784	0.01606
vsource[38]	0.03453	0.03986	0.05531	-0.09454	-0.15091
vsource[39]	-0.01319	-0.02330	0.10926	0.01625	0.08168
vsource[40]	0.01256	0.00027	-0.00757	0.00500	0.00466
vsource[41]	0.00103	-0.12218	0.05007	-0.03393	0.08424
vsource[42]	-0.10440	0.05686	0.05354	-0.07094	0.06555
vsource[43]	-0.00607	-0.03504	0.07928	-0.24311	-0.16625
vsource[44]	0.02401	-0.12169	-0.03863	0.33831	0.08239
vsource[45]	0.01588	0.07843	0.01200	-0.00299	0.04841
vsource[46]	0.01298	-0.00209	-0.01397	-0.00129	0.00237
vsource[47]	0.03644	0.00021	-0.03267	0.01822	0.02633
vsource[48]	0.04843	0.00800	-0.01985	-0.03666	0.00475
vsource[49]	0.00103	0.00274	-0.02263	0.05518	0.00626
vsource[50]	0.02948	0.00566	0.03397	-0.00822	-0.00121
vsource[51]	0.00424	0.00935	-0.00489	-0.00758	0.00378
vsource[52]	0.04895	-0.03666	-0.07267	0.09535	-0.02881
vsource[53]	-0.00252	-0.00461	0.00404	-0.00960	0.01956
vsource[54]	0.07465	-0.09834	-0.03488	-0.09258	0.17421
vsource[55]	0.02051	-0.00324	0.11754	0.09613	-0.03557
vsource[56]	0.00864	-0.00295	-0.01506	0.00359	0.00338
vsource[57]	-0.01277	-0.01189	0.09888	0.14572	-0.03735
vsource[58]	0.03139	-0.00041	-0.02960	-0.01241	0.08979
vsource[59]	0.00865	0.00288	-0.04360	0.02346	0.00913
vsource[60]	-0.01766	-0.02605	0.02992	0.00387	0.01015
vsource[61]	0.05388	0.06652	-0.10053	-0.04706	0.02660
vsource[62]	-0.16553	-0.07490	0.07378	0.03005	0.05483
vsource[63]	0.00671	-0.00622	-0.04960	-0.02595	-0.04217
vsource[64]	-0.00061	-0.00990	-0.02383	0.00408	-0.02147
vsource[65]	-0.06918	-0.41015	-0.31227	0.04640	0.08656
vsource[66]	0.18671	0.23736	-0.03242	-0.14391	0.33481
vsource[67]	0.02632	-0.07329	-0.03016	-0.05330	-0.06910

vsource[68]	-0.05061	0.15374	0.43027	0.24791	0.22328
vsource[69]	0.29684	-0.05082	-0.18065	0.20999	-0.30140
vsource[70]	0.12464	0.01418	-0.09899	-0.11312	0.10003
vsource[71]	0.00359	0.00216	-0.03021	0.02051	-0.01008
vsource[72]	0.01965	-0.00170	-0.02107	0.00402	0.00409
vsource[73]	-0.00204	-0.17034	0.11004	-0.21162	0.08479
vsource[74]	-0.54862	-0.00589	-0.26586	0.23467	-0.02711
vsource[75]	0.02770	0.01507	0.01163	-0.03641	-0.01307
vsource[76]	-0.05078	-0.08720	-0.16182	0.00420	-0.01609
vsource[77]	0.00882	-0.01611	-0.03578	0.04834	-0.02714
vsource[78]	0.36117	0.29567	-0.06625	0.21689	0.14602
vsource[79]	0.15934	-0.35379	0.41690	0.01397	-0.07317
vsource[80]	0.26877	-0.24431	0.14766	-0.15252	-0.28770
vsource[81]	0.01941	-0.06423	0.00267	0.31293	-0.03351
vsource[82]	0.03370	-0.01590	-0.09313	0.12806	0.00171
vsource[83]	0.05517	0.08873	-0.08087	0.19940	-0.21834
vsource[84]	0.10014	-0.04337	-0.11739	-0.14736	-0.04141
vsource[85]	0.25769	-0.45409	-0.07267	0.20898	0.37609
vsource[86]	0.18474	0.16469	-0.05516	0.06098	0.01985
vsource[87]	0.00793	-0.00438	-0.00951	0.07053	0.02091
vsource[88]	0.04689	-0.10376	-0.16617	-0.05171	-0.20839
vsource[89]	0.06731	0.00734	0.08512	-0.00730	-0.00559
vsource[90]	0.11176	0.13956	0.11107	0.32050	-0.09664
vsource[91]	0.03540	-0.02538	0.01623	0.05084	-0.02705
vsource[92]	0.00382	0.00474	-0.01373	0.01469	-0.00371
vsource[93]	0.09904	-0.09935	-0.11266	-0.02325	-0.04444
vsource[94]	0.01031	-0.01646	-0.03345	0.00860	-0.03575
vsource[95]	0.00220	-0.00130	-0.01121	-0.00567	-0.00915
vsource[96]	0.01461	0.01121	-0.02338	0.03614	0.03611

2) Nov 98 EcoMicroPlates™

PCA Scores					
Plot	First Component	Second Component	Third Component	Fourth Component	Fifth Component
FYM (rep 1)	0.4585	0.1970	0.3253	0.4449	0.0599
FYM (rep 2)	0.4296	0.1997	0.3059	-0.4483	0.0854
FYM (rep 3)	0.0059	0.2745	-0.0868	-0.0243	-0.3346
Sewage (rep 1)	0.1771	0.3076	-0.5893	0.0191	0.1427
Sewage (rep 2)	0.2147	-0.9935	-0.0786	-0.0025	-0.0183
Sewage (rep 3)	-1.2859	0.0147	0.1235	0.0110	0.0650

Latent Vectors (Loadings)

	1	2	3	4	5
vsource[2]	0.00264	-0.00917	0.00052	-0.00712	-0.00959
vsource[3]	0.01511	0.07376	-0.40452	0.19773	0.19712
vsource[4]	-0.09699	0.01586	-0.44314	-0.11504	-0.12495
vsource[5]	0.06296	0.04534	0.17935	-0.37629	0.19942
vsource[6]	0.03182	0.05430	0.04419	0.02193	0.16554
vsource[7]	0.17557	0.09452	0.14948	0.06938	-0.44428
vsource[8]	0.10812	-0.76967	0.21538	-0.04730	-0.15248
vsource[9]	0.03155	0.10504	0.15094	0.04028	-0.17547
vsource[10]	0.02114	0.01672	0.05432	-0.04763	0.04231
vsource[11]	0.00000	0.00000	0.00000	0.00000	0.00000
vsource[12]	0.01403	0.00440	0.01084	-0.02486	-0.06289
vsource[13]	0.08578	0.03185	-0.03211	0.01564	-0.04934
vsource[14]	0.03623	0.10461	0.09620	-0.18941	0.09618
vsource[15]	-0.05965	-0.31495	-0.08050	0.05171	0.34340
vsource[16]	-0.79768	-0.01134	0.18013	0.03342	-0.34145
vsource[17]	0.01359	-0.01896	0.00688	-0.01379	-0.06839
vsource[18]	0.01471	0.01787	-0.21578	0.22503	0.00771
vsource[19]	0.16346	0.16416	0.45643	0.10523	-0.03474
vsource[20]	0.00464	0.00100	0.00479	-0.01722	-0.04541
vsource[21]	0.07063	-0.45444	-0.10098	0.00183	-0.04170
vsource[22]	0.01047	0.00120	-0.06657	0.10714	-0.08996
vsource[23]	0.13168	0.04862	0.31908	0.63614	0.19820
vsource[24]	0.00034	-0.01339	-0.00439	-0.01284	-0.05921
vsource[25]	-0.45558	0.00527	0.15540	0.01631	0.32935

vsource[26]	0.07294	0.05513	0.18564	-0.40051	0.18932
vsource[27]	0.00572	-0.00481	0.00056	-0.00301	-0.01971
vsource[28]	-0.00071	0.02992	0.13237	-0.25633	0.13782
vsource[29]	0.00114	-0.00544	-0.00519	0.00014	0.00287
vsource[30]	0.00286	0.02421	-0.03941	0.08145	-0.03903
vsource[31]	-0.15797	-0.15157	0.10999	0.20599	0.27317
vsource[32]	-0.07257	0.03210	-0.02895	-0.00930	0.27280

3) **Mar 99 BIOLOG™**

PCA Scores					
Plot	First Component	Second Component	Third Component	Fourth Component	Fifth Component
FYM (rep 1)	1.081	2.075	3.511	4.593	-4.199
FYM (rep 2)	-8.843	-4.473	-1.626	-2.138	1.472
FYM (rep 3)	7.606	-3.640	3.458	-3.282	1.504
Sewage (rep 1)	-2.157	6.726	1.663	-3.937	-1.951
Sewage (rep 2)	0.051	3.413	0.526	4.241	4.850
Sewage (rep 3)	4.424	0.050	-7.533	0.523	-1.676

Latent Vectors (Loadings)

	1	2	3	4	5
vsource[2]	-0.04543	0.15712	0.12152	-0.10872	-0.06561
vsource[3]	-0.15260	-0.07733	-0.05071	-0.06050	0.06129
vsource[4]	-0.10637	0.14958	-0.08343	-0.04818	0.07501
vsource[5]	-0.08988	0.09854	0.11114	0.11648	0.12028
vsource[6]	-0.11716	-0.16335	0.01272	0.04964	0.03837
vsource[7]	0.00076	0.16176	-0.03331	0.10049	0.18170
vsource[8]	0.14430	0.05095	0.04359	0.13157	0.01602
vsource[9]	-0.04955	0.13867	-0.06903	-0.15914	-0.10687
vsource[10]	-0.14938	0.09333	-0.01979	-0.08424	-0.02860
vsource[11]	0.00801	0.17862	0.14131	0.05362	-0.02860
vsource[12]	-0.14282	0.09269	0.07656	-0.06768	0.03544
vsource[13]	0.06712	0.00129	-0.21279	0.01824	-0.07867
vsource[14]	-0.02960	-0.04648	0.12986	-0.18792	0.11689
vsource[15]	-0.17027	-0.00680	0.02562	0.05608	-0.02689
vsource[16]	-0.15138	0.09048	-0.06098	0.03327	0.04932
vsource[17]	-0.11384	0.14614	-0.02434	0.10775	-0.02152
vsource[18]	-0.07780	0.09923	-0.17360	-0.03510	0.08681
vsource[19]	0.09237	-0.01757	0.02938	-0.13511	0.20702
vsource[20]	-0.06583	-0.05722	0.06196	0.18768	-0.15069
vsource[21]	-0.12124	0.02403	-0.06701	0.16999	0.05439
vsource[22]	-0.02335	0.05907	0.09491	0.18543	0.16303
vsource[23]	-0.13576	0.00087	-0.00454	-0.14566	-0.10108
vsource[24]	-0.11972	0.12228	0.00799	0.11363	-0.08324
vsource[25]	-0.11470	0.15503	0.06282	-0.04782	-0.04453
vsource[26]	-0.01883	-0.09180	-0.18417	0.05639	-0.13720

vsource[27]	-0.16900	0.04111	0.04197	0.02815	0.02014
vsource[28]	-0.09004	0.07938	-0.07998	0.04141	0.21537
vsource[29]	-0.16417	-0.01665	-0.07488	-0.03659	0.02861
vsource[30]	-0.00816	-0.00116	-0.23548	-0.02897	-0.04834
vsource[31]	-0.15537	0.00571	0.02481	-0.02582	-0.13841
vsource[32]	-0.14818	-0.11890	-0.00254	0.02959	0.03334
vsource[33]	-0.16225	-0.00610	-0.08598	0.03202	-0.02533
vsource[34]	0.09475	-0.15333	0.03398	-0.11568	0.07460
vsource[35]	-0.02234	0.16192	0.01886	-0.17795	-0.06018
vsource[36]	0.15447	0.08248	0.06213	0.01805	-0.05325
vsource[37]	0.06712	0.00129	-0.21279	0.01824	-0.07867
vsource[38]	-0.09436	-0.13431	0.12667	-0.07098	0.04669
vsource[39]	0.11519	0.08910	0.15468	0.00073	0.02529
vsource[40]	-0.03820	0.00380	-0.04457	0.24175	0.09667
vsource[41]	-0.10689	0.04444	0.02868	-0.18926	0.08565
vsource[42]	-0.15002	0.03250	-0.11365	-0.04535	-0.00674
vsource[43]	-0.05022	-0.05969	0.20220	-0.09582	0.03572
vsource[44]	-0.06609	-0.21011	-0.00969	0.01718	0.04298
vsource[45]	-0.16921	-0.03109	0.05216	-0.01192	-0.02786
vsource[46]	0.00078	0.08851	0.01487	0.14798	0.22759
vsource[47]	-0.15124	0.01880	-0.07055	-0.03930	0.11860
vsource[48]	-0.14610	-0.01595	0.07130	0.11884	0.04321
vsource[49]	0.04420	-0.00342	-0.21989	0.02443	-0.09275
vsource[50]	0.00000	0.00000	0.00000	0.00000	0.00000
vsource[51]	0.00000	0.00000	0.00000	0.00000	0.00000
vsource[52]	-0.14562	0.09189	0.02870	-0.00539	-0.11628
vsource[53]	-0.13417	-0.11600	-0.04592	-0.07459	0.06910
vsource[54]	-0.07317	-0.04081	-0.05688	-0.03934	0.26223
vsource[55]	0.06613	0.08765	0.20192	-0.02052	0.00839
vsource[56]	-0.12136	-0.04871	-0.15664	0.00601	0.07169
vsource[57]	-0.05826	0.08311	0.19072	0.00111	0.11123
vsource[58]	-0.09433	0.18524	0.05747	-0.00503	0.02574
vsource[59]	-0.14605	0.12248	-0.00080	0.00633	-0.05112
vsource[60]	-0.07203	-0.06184	0.20273	-0.04993	-0.03453
vsource[61]	0.03839	-0.20532	0.04251	-0.09423	0.00108
vsource[62]	-0.09115	-0.16566	0.01486	0.04591	-0.13128
vsource[63]	-0.14930	0.09372	0.02689	-0.00049	-0.09875
vsource[64]	-0.04117	-0.04267	-0.06921	-0.20547	0.15004
vsource[65]	0.00233	-0.01951	0.14327	-0.20919	-0.04612
vsource[66]	0.14115	0.11184	-0.05352	-0.01684	-0.07996
vsource[67]	-0.00448	-0.09896	0.04316	-0.15167	-0.20969
vsource[68]	0.14753	-0.05138	-0.09388	-0.01181	-0.09415
vsource[69]	0.10238	0.18160	-0.00929	-0.00713	0.05901
vsource[70]	0.09238	0.16729	-0.00487	0.09349	-0.08342
vsource[71]	-0.13463	0.03955	0.00659	0.06156	-0.17815
vsource[72]	-0.15208	-0.07086	0.07289	-0.06026	-0.03816
vsource[73]	0.04984	-0.18296	0.02869	0.13846	0.01958
vsource[74]	0.01251	0.20753	-0.06210	0.05266	-0.08508

vsource[75]	0.01879	-0.03366	-0.10282	-0.23528	-0.02501
vsource[76]	-0.07182	-0.03063	-0.14884	0.12960	-0.13693
vsource[77]	-0.11291	0.01467	-0.08629	-0.14464	0.12435
vsource[78]	0.02230	-0.19222	0.11997	0.04776	-0.02043
vsource[79]	-0.07174	-0.05278	0.17329	0.05210	-0.14594
vsource[80]	-0.03273	0.17442	0.04697	-0.13737	-0.09155
vsource[81]	0.10423	0.04948	-0.03673	-0.04391	0.23006
vsource[82]	0.06148	0.19654	0.08336	-0.03194	-0.03584
vsource[83]	0.01344	0.04769	-0.15835	-0.14961	0.13912
vsource[84]	0.05441	-0.11042	-0.18697	0.06949	0.00163
vsource[85]	-0.00216	-0.03458	0.17468	0.07569	0.18768
vsource[86]	-0.04338	-0.06134	0.18684	-0.13468	0.02789
vsource[87]	-0.03111	-0.14022	0.14157	0.12998	0.03465
vsource[88]	-0.10733	-0.11549	-0.13104	-0.06730	0.03763
vsource[89]	0.01649	0.09060	-0.03461	0.15523	0.21378
vsource[90]	-0.12765	-0.09873	0.03616	0.13724	-0.01583
vsource[91]	-0.08077	-0.00082	0.11619	0.06424	-0.21826
vsource[92]	-0.16399	0.03991	-0.05873	-0.02925	-0.05338
vsource[93]	-0.16879	-0.01098	-0.04774	0.05122	-0.01202
vsource[94]	-0.08564	-0.14193	0.03402	0.13383	0.10319
vsource[95]	-0.14658	0.01230	-0.06554	-0.08028	0.11516
vsource[96]	-0.07927	0.04843	0.03031	0.22737	-0.03241

4) **Mar 99 EcoMicroPlates™(a)**

PCA Scores					
Plot	First Component	Second Component	Third Component	Fourth Component	Fifth Component
FYM (rep 1)	5.1972	-0.6656	1.3124	0.8383	-1.4666
FYM (rep 2)	-0.1332	5.8942	0.2082	-0.8592	0.4273
FYM (rep 3)	1.7410	-1.9698	-2.7367	-1.6823	1.8925
Sewage (rep 1)	-1.6300	-0.4922	0.2014	3.3704	1.7105
Sewage (rep 2)	-2.7644	-0.5725	-2.1082	0.1907	-2.8056
Sewage (rep 3)	-2.4105	-2.1941	3.1229	-1.8579	0.2420

Latent Vectors (Loadings)

	1	2	3	4	5
vsource[2]	-0.05964	0.28264	0.03290	0.16383	0.21097
vsource[3]	-0.13466	0.04970	0.20164	0.33000	-0.24612
vsource[4]	-0.03903	-0.32370	-0.04253	0.04451	0.10695
vsource[5]	0.25468	0.19825	0.07494	-0.05997	-0.05867
vsource[6]	-0.21647	-0.02975	0.18991	0.21276	0.25429
vsource[7]	0.18794	-0.00646	0.20169	-0.01932	0.37866
vsource[8]	-0.04091	-0.32440	-0.06731	0.08207	0.02428
vsource[9]	0.01217	-0.32780	0.00201	0.07985	-0.07525
vsource[10]	-0.15177	-0.11912	0.32527	-0.20431	0.03637
vsource[11]	0.00000	0.00000	0.00000	0.00000	0.00000
vsource[12]	-0.23497	-0.07700	0.21585	-0.23399	0.05241
vsource[13]	0.12514	-0.13090	0.33770	-0.20841	0.01848
vsource[14]	0.26747	-0.12426	-0.01077	0.13475	0.19850
vsource[15]	0.29977	-0.12405	0.04308	-0.06651	-0.04830
vsource[16]	0.10299	-0.04791	-0.24660	-0.39215	0.04848
vsource[17]	-0.25050	0.20394	-0.10358	-0.00967	-0.00470
vsource[18]	0.01959	-0.07575	0.34663	0.09823	0.32025
vsource[19]	-0.13832	-0.23263	-0.11711	0.19770	0.19563
vsource[20]	-0.26682	0.03704	-0.16724	0.07251	-0.23097
vsource[21]	-0.06782	0.31506	0.04711	0.05462	0.12602
vsource[22]	0.18539	-0.03367	0.11235	-0.34305	-0.22063
vsource[23]	0.27574	-0.03668	0.13683	0.10740	-0.21421
vsource[24]	-0.26705	-0.17511	-0.02281	-0.12769	-0.03491
vsource[25]	-0.20487	-0.14504	-0.02828	-0.20758	-0.27677
vsource[26]	-0.13172	-0.01278	0.33661	-0.27783	0.05678
vsource[27]	0.04971	0.28335	-0.09775	-0.19553	0.14687

vsource[28]	0.25978	-0.04286	0.14499	0.21709	-0.15592
vsource[29]	0.09237	-0.10855	-0.28534	-0.21554	0.27641
vsource[30]	-0.22002	-0.22574	-0.03068	-0.01721	0.16781
vsource[31]	-0.12188	0.26831	0.00157	-0.16758	0.18554
vsource[32]	-0.14546	0.12764	0.32655	-0.02404	-0.21604

5) **Mar 99 EcoMicroPlates™(b)**

PCA Scores					
Plot	First Component	Second Component	Third Component	Fourth Component	Fifth Component
FYM (rep 1)	-1.7622	-0.1466	0.7350	3.6157	1.7870
FYM (rep 2)	2.4549	4.9971	-0.3969	- 0.3121	0.0226
FYM (rep 3)	-2.0664	-0.6826	- 4.0062	0.5097	-1.5068
Sewage (rep 1)	-1.1336	-0.3697	3.7072	0.0192	-2.0545
Sewage (rep 2)	-2.3404	-0.5645	0.2250	-3.7272	1.5082
Sewage (rep 3)	4.8477	-3.2338	-0.2640	- 0.1053	0.2436

Latent Vectors (Loadings)

	1	2	3	4	5
vsource[2]	0.27195	-0.21778	-0.02113	-0.00943	0.04964
vsource[3]	0.27360	-0.08777	-0.10827	0.06982	-0.28248
vsource[4]	0.30912	-0.03487	0.02886	0.16311	-0.05036
vsource[5]	0.26917	0.19346	-0.00080	-0.12508	0.06276
vsource[6]	0.12588	-0.01775	-0.09693	-0.32186	0.31299
vsource[7]	0.07103	-0.00661	-0.29184	-0.28000	0.04835
vsource[8]	-0.05157	0.23452	0.18932	-0.25529	-0.02021
vsource[9]	-0.21042	-0.28299	-0.06305	-0.00731	0.05057
vsource[10]	-0.09421	0.15020	-0.30863	-0.16356	-0.10981
vsource[11]	0.00000	0.00000	0.00000	0.00000	0.00000
vsource[12]	0.26224	-0.11796	-0.20093	0.04519	-0.12876
vsource[13]	0.14960	0.00038	0.33079	-0.15692	-0.00465
vsource[14]	0.02736	-0.00742	-0.24705	-0.22247	-0.38095
vsource[15]	-0.12145	-0.06497	-0.09410	-0.37550	-0.08023
vsource[16]	-0.19492	-0.26024	0.10706	-0.12010	0.10510
vsource[17]	0.12706	-0.15107	-0.24781	-0.17278	-0.25296
vsource[18]	0.10268	-0.12333	0.30681	-0.18646	0.11342
vsource[19]	0.00989	0.15778	0.25417	-0.26039	-0.14691
vsource[20]	-0.13888	0.01290	0.22357	-0.30943	-0.01930
vsource[21]	0.15477	0.32547	0.02251	-0.02955	-0.07179
vsource[22]	0.20325	-0.22439	0.18166	0.04975	-0.15452
vsource[23]	0.13771	0.33654	-0.03176	-0.02797	0.00460
vsource[24]	0.27659	-0.06448	0.20710	0.00912	0.12675
vsource[25]	0.26626	-0.10835	0.00108	0.13176	0.28885
vsource[26]	0.04897	-0.12069	0.00106	-0.33474	0.32924
vsource[27]	0.14187	-0.12547	0.27174	-0.09454	-0.29469
vsource[28]	-0.06359	-0.02490	0.29665	0.00172	-0.41870

vsource[29]	0.27973	0.20241	-0.04183	-0.03183	0.03146
vsource[30]	0.03432	-0.30193	-0.08887	-0.21188	-0.11686
vsource[31]	0.11887	0.32973	-0.04059	-0.11686	0.02047
vsource[32]	0.25299	-0.20495	-0.07441	-0.13010	0.06139

Procrutes Rotation

A) Ecoplates 98 v' BIOLOG GN1 98

Sums of Squares eco88vb962;***

Fitted Configuration	11.6809
Residual	8.6063
---	---
Fixed Configuration	3.0746

B) Ecoplates 98 v' Ecoplates 99 (a)

*** Sums of Squares eco98vb99***

Fitted Configuration	5.8733
Residual	2.2788
---	---
Fixed Configuration	8.1520

C) Ecoplates 99 (a) v' Ecoplates 99 (b)

*** Sums of Squares ecorep99***

Fitted Configuration	0.9006
Residual	2.8108
---	---
Fixed Configuration	3.7114

D) Ecoplates 98 v' Ecoplates 99(b)

*** Sums of Squares *eco992v98**

Fitted Configuration	1.2839
Residual	2.7475
---	---
Fixed Configuration	4.0314

e) EcoPlate v's EcoPlate 98

*** Sums of Squares **

Fitted Configuration	3.1846
Residual	4.9675
---	---
Fixed Configuration	8.1520

Appendix C

16S Restriction Profiles

The restricted products generated a distinct restriction pattern, these species-specific patterns (mostly) reflect the conserved character of the rRNA genes (size in bp) estimated from λ *hind* III digest.

(a) Restriction Patterns of Reference Strains (9.3.2)

i) 16S rRNA Universal Profiles (see 9.3.2.1)

	4002	**4003**	**4004**	**4008**	**4009**	**4010**	**10527**	**1007**
***Alu* I**								
3	3	3	3	3	3	3	5	3
6	6	6	6	6	5	6	6	6
7	7	7	7	7	6	7	7	7
	9	8	8		7	8	8	8
175	97	100	100	170	170	104	104	104
210	166	166	166	215	220	170	175	175
420	215	220	220	420	287	226	220	220
	420	430	430		390	430	280	280
***Hin* FI**								
					3			3
1	1	1	1	1	6	1	81	3
7		7	7	7	7	8		8
130		130	130	130	130	113.3	113	113
837	837	806	806	837	229	837	837	502
					517			532
***RSA* I**								
3	2	3	3	3		3	3	4
4		4	4	4	5	4	4	5
7	7	7	7	7	7	7	7	5
128	140	145	145	145	134	134	128	123
380	391	402	391	391	245	379.6	369	230
519	551	535	519	534		519	504	277
								359

ii) 16S rRNA Specific Profiles (see 9.3.2.2)

4000	**4002**	**4003**	**4004**	**4008**	**4010**	**10527**	**1007**
***Alu* I**							
20	20	20	20	20	20	40	40
50	50	50	50	50	50	50	50
***Hin* FI**							
10	10	10	10	10	10	10	10
			20	20			30
			30	20			
			30	30			30
RSA I							
30	30	50	30	30	30	30	40
50	50	50	50	50	50	50	60

(b) **Restriction Patterns of Field Isolates (9.4.1)**

16S Ribosomal Restriction Digest Pattern using the *Pseudomonas* Specific Primers on the Nov Field Isolates.

***RSA* I**										
fn 1	649	374	338		sn 1	0	0	0		
fn 2	1051	649	392		sn 2	554	321	304		
fn 3	649	240			sn 7	554	321	304		
fn 4	649	392			sn 8	554	321	272		
fn 6	624	374	338		sn 9	554	304			
fn 8	577	321			sn 11	739	473			
fn 11	1013	676	449		sn 12	1012	709	450		
fn 12	751	573			sn 13	0	0	0		
fn 13	751	449			sn 14	770				
fn 14		449			sn 15	0	0	0		
fn 13	751	606	166		sn 17	13	18	23		
fn 16	713	420			sn 17	995	678	410		
fn 17	623	551			sn 18	995	678	428		
fn 19	701	605			sn 19	703	428			
fn 20	681	419			sn 20	703	465			
fn 21	662	274			sn 21	678	359	294		
fn 22	359	122	76		sn 22	654	836	780	728	234
fn 23	701	359			sn 23	666	386	328	300	222
fn 24	710	381	248		sn 24	666	342	300		
fn 27	710	335			sn 25	666	401	356	328	
fn 28	710	291			sn 29	988	666	386	342	
fn 29	742	430			sn 30	1018	988	666	386	342
fn 30	742	430	207		sn 31	1018	988	666	401	
fn 2	710	430	248							

Cfo I								
fn 1	304	272	166		sn 1	0	0	0
fn 2	304	272	152		sn 2	240	195	110
fn 3	321	272	166		sn 7	272	225	124
fn 4	288	256	152		sn 8	240	195	97
fn 6	272	240	138		sn 9	240	180	87
fn 8	0	0			sn 11	340		
fn 11	364	311	213		sn 12	361	184	
fn 12	392	337	236		sn 13	0	0	
fn 13	392	324	0		sn 14	382		
fn 14	392	338	236		sn 15	0	0	
fn 13	364	311	212		sn 18	248	129	
fn 16	286	166			sn 17	342	294	192
fn 17	330	274			sn 18	359	310	206
fn 19	330	301	195		sn 19	326	294	206
fn 20	344	301	221		sn 20	326	294	206
fn 21	359	76			sn 21	326	279	151
fn 22	330	195			sn 22	294	249	151
fn 23	330				sn 23	314	260	162
fn 24	335	290	167		sn 24	314	260	174
fn 27	335	167			sn 25	314	273	174
fn 28	313	207			sn 29	314	260	162
fn 29	381	335			sn 30	314	260	174
fn 30	381	335			sn 31	314	260	151
fn 2	290	168	22					

***Hae* III**											
fn 8	600	180			sn 1	532	22				
fn 1	624	195			sn 2	554	138				
fn 2	624	195			sn 9	577	151				
fn 6	624	195			sn 7	577	166				
fn 3	648	225			sn 8	600	166				
fn 4	648	209			sn 12	739	427				
fn 11	965	676	236		sn 11	739					
fn 27	678	207			sn 13	0	0				
fn 17	661	221			sn 14	770					
fn 19	661	247			sn 25	666	356	314	4		
fn 20	681	260			sn 31	1007	646	342			
fn 21	681	274			sn 15	1092					
fn 22	681	260			sn 17	678	249				
fn 23	701	247			sn 16	647	290	167			
fn 24	709	248			sn 18	702	249				
fn 12	713	261			sn 19	702	249	234	100	6	
fn 28	709	429			sn 20	702	249	206	76	6	
fn 16	751	286			sn 30	666	328	222			
fn 13	751	298			sn 21	678	358	310	222	85	
fn 14	771	311			sn 22	631	358	326	222	85	4
fn 29	742	248			sn 23	666	356	300	235	106	
fn 30	742	248			sn 24	666	342	300	117	97	
fn 2	709	248	381		sn 29	666	356	314	300	198	
fn 13	732	298	122								