Woodcock, Stephen Mark (2007) *Modelling the assembly and structure of microbial communities with applications to waste treatment strategies.*

PhD thesis

Modelling the Assembly and Structure of Microbial Communities with Applications to Waste Treatment Strategies

Stephen Mark Woodcock

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy in Civil Engineering

University of Glasgow
Department of Civil Engineering

July 2007

To my long-suffering sidekick...

Abstract

Increasing global population sizes and industrialisation mean that the need for efficient, reliable and cheap wastewater treatment strategies grows ever greater. The majority of current strategies rely on naturally occurring microbial communities to metabolise pollutants. These have developed slowly through many years of empirical research. Despite our reliance on microbial communities, our understanding of how they assemble and change through time remains relatively poorly understood. The reason for this is that the communities have until recently been very difficult to observe *in situ.* There are, however, a number of new techniques in the field of molecular microbial ecology which are affording engineers new insight into the composition of these communities.

The central premise of this thesis is that combining models in theoretical ecology with our newfound ability to observe and measure microbial systems will allow for a suite of laws to be developed which can describe and predict the assembly and structure of microbial communities. This would allow environmental engineers to modify and improve the design and efficiency of wastewater treatment systems.

It is demonstrated here that there is significant evidence that microbial community assembly is at least partly a random process. A simple Neutral Community Model (NCM) is shown to replicate much of the variability observed in real systems as diverse as the waste water treatment plants, estuaries and the human lung. This is in contrast to the prevailing view in microbial ecology that community composition is shaped by deterministic processes. Molecular methods in microbial ecology yield very small, sometimes biased, samples from what are ostensibly very large communities. It is demonstrated, using published literature on taxa-area relationships for microorganisms that sampling effects have the capacity to significantly distort the true, underlying ecological patterns. In doing so, a potential reconciliation is offered between some seemingly contradictory published reports on the nature of taxa-area relationships for microorganisms.

The effects of sampling are built directly into the NCM, which allows the model to be tested using the data which are typically collected by microbial ecologists. The model is calibrated using the taxa-abundance distribution observed in a small waterborne bacterial community housed in a bark lined tree hole in a beech tree. Using these parameters, unchanged, it is shown that the model can predict the taxa-abundance distributions and taxa-volume relationship observed in 26 other beech tree communities whose sizes span three orders of magnitude. This represents the strongest test, so far, for any biological community, microbial or otherwise, that NCMs provide a useful tool in predicting community composition.

Table of Contents

List of Figures

List of Tables

Table of Symbols

N_i	The number of individuals of species i in the local community.
N_T	The number of individuals in total in the local community ($\sum_i N_i = N_T$).
N_S	The number of individuals in total in a sample.
S_T	The number of species in the local community.
S_M	The number of species in the metacommunity.
θ	The fundamental biodiversity number, a scaling parameter for the logseries distribution
p_i	The relative abundance of species i in the metacommunity.
x_i	The relative abundance of species i in the local community ($x_i = \frac{N_i}{N_T}$).
y_i	The relative abundance of species i in a sample.

Acknowledgements

First and foremost, I would like to express my deepest and most sincere thanks to Professor Bill Sloan. Over the past three and a half years, he has not only offered the greatest of support and help with my research but has also become a valued friend. I feel very fortunate to have worked alongside him, and I am deeply grateful for all his efforts during my time at the University of Glasgow.

The work undertaken for this thesis was done in conjunction with researchers at the University of Newcastle-upon-Tyne. It would take too long to thank everyone who has been of assistance individually, but I would like to express particular thanks to Professor Tom Curtis, with whom I have enjoyed a fruitful and often entertaining collaboration. Additionally, I would like to thank Dr. Mary Lunn at St. Hugh's College, Oxford for her support and academic guidance that have now stretched back almost eight years.

Finally, I would like to thank Polly Putnam for her patience and understanding and for the countless long telephone calls which she has had during the duration of my time in Glasgow. She has offered her continuous support throughout my research and has made the long and often difficult process of writing this thesis much more enjoyable than it may otherwise have been.

Author's Declaration

I declare that this thesis is a record of the original work carried out solely by myself in the Department of Civil Engineering at the University of Glasgow, United Kingdom, during the period of October 2003 to May 2007. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright acts. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis. This thesis has not been presented elsewhere in consideration for a higher degree.

Stephen Mark Woodcock

July 2007

1 Introduction

With the global increase of population sizes and expansion of the manufacturing sector, the need for treatment of waste, both industrial and domestic has never been greater. In highly industrialised and in developing nations alike, mankind's basic needs are the same. Our vital requirements of sanitary water and soils which are fertile and free from any dangerous contaminants are universal. Yet, at the present time, we are unable to provide such necessities to almost half of the world's population. The challenge of devising waste treatment strategies which are efficient, in terms of both time and cost, is one of the greatest facing environmental engineers today (Newman & Mouritz 1996).

Since early in the last century, the basic concept behind most waste treatment systems has changed very little (Rittmann & McCarty 2001). Employment of biodegradation strategies, which rely on naturally forming microbial communities to remove pollutants, is still very much standard practice. If microorganisms which metabolise chemicals within the waste can be encouraged to flourish, then the mass of pollutants will diminish. For example, communities of ammonia oxidising bacteria (AOB) are encouraged in the treatment of domestic waste rich in ammonia.

Despite being a very well established technology, certainly in terms of the number of years it has been employed, many aspects of it are still relatively poorly understood. Our knowledge of the underlying ecology of the microbes, and thus the very foundation of the strategy, remains somewhat limited. This means that our efforts to design bioreactors based on chemical thermodynamic principles (Sorochenko 2001) without reference to the ecology of the microbial catalysts are only partially successful. It is only due to the wealth of empirical data gathered from years of research and application that we are able to engineer systems such as those which treat our drinking water every day. However, this reliance solely upon experience is less than ideal, and even the most successful waste treatment plants suffer occasional loss of function (Curtis *et al.* 2003).

In the mindset of empiricism, these failures are baffling. Why should the same strategy work on certain occasions or in certain locations, and yet be unsuccessful in others? Also, stepping beyond the bounds of established systems is risky. It is a brave engineer who radically alters the design of a system to something unprecedented, for fear of upsetting the tried and tested procedures. This necessary conservatism in design severely restricts the spread of the technologies to some of the geographical areas with socio-economic and health problems, where they are most desperately needed.

Many developing nations currently lack the infrastructure to support large scale wastewater treatment plants, which experience tells us tend to be more reliable. In addition to this, re-establishing function in a failing system is not a cheap or simple procedure; often huge quantities of waste must be drained away, or have highly expensive agents added. As with all empirically derived strategies, these techniques tend to work, but their success in any given instance is certainly not assured. For regions which lack the finances to maintain such a costly operation, plants built to current design are fundamentally not a feasible option (Saldinger 1992). The development of a new generation of highly efficient waste treatment systems may not only bring economic benefits for the developed world, but also save the lives of those who presently lack a dependable source of sanitary drinking water.

Other engineering disciplines have long since been liberated from such reliance on empiricism. Ever since Newtonian mechanics were first applied, structural engineers have had the tools to calculate whether or not radically new structures will be functional. Environmental engineers working with waste treatment have yet to be afforded such peace of mind.

Until relatively recently, it was almost impossible to observe microbial communities *in situ*. However, the last twenty years have seen the emergence of a number of new technologies, which afford engineers glimpses like never before into the microbial world. The composition of communities can now be characterised using molecular methods by, for example, examining distinguishing features of the 16S rRNA sequences of microbes' functional genes (Lane 1991). Via such laboratory tools, we can gain unprecedented levels of information about the structure and diversity of microbial communities. One can only speculate with excitement as to what further tools will be developed during the next twenty years. However, despite the wealth of new information being gathered in laboratories, there has as yet been little impact upon engineered biological systems and waste treatment strategies.

All these new developments in the laboratory may well stand us on the brink of a complete revolution in the design of engineered biological systems. It is now centuries since simple observations were used to formulate universal laws in classical physics, which were later applied by engineers. The challenge currently facing environmental engineers is to derive a similar suite of 'laws' dictating the assembly and composition of communities of microorganisms from our newfound abilities to observe such systems.

That said, entire wastewater treatment plants are hugely complex systems, typically consisting of the order of 10^{18} individual microbes of many different taxa (Whitman *et al.* 1998). Even the most wildly optimistic engineer cannot expect that we will ever be able to classify each and every organism within such a community. Instead, we must always work from samples which contain a tiny fraction of the total population. To extrapolate information from such subsections of a system is not always a trivial task, and we must be aware as to which observed phenomena are likely indicative of the community as a whole, and which are merely artefacts of the sampling procedures employed. These limitations and problems have, as yet, meant that despite the revolution of molecular methods, there is little evidence of major impact in engineering practice.

To determine rules of community assembly for microorganisms necessarily takes the majority of the research presented in this thesis into realms not traditionally considered by environmental engineers. The focus here is very much upon developing fundamental theories in microbial ecology rather than answering any one practical problem or enhancing a given reactor design.

The central premise of this thesis is that, even only working from sample data, environmental engineers should still aim to determine a set of rules governing community assembly and structure, just as structural engineers did many decades ago. The establishment of such theorems in this field, combined with our sample data, should permit the derivation of simple parameterisable models of microbial communities. These models could then be applied to inform a new generation of waste treatment systems, ones more reliable and more easily manipulated than the legacy of empiricism currently allows. The benefits to some are purely financial, in the form of lower utility bills. To others, our ability to develop such waste treatment strategies is, quite literally, a matter of life and death.

The research presented in this thesis forms the basis of five papers published in peer-reviewed journals; Environmental Microbiology (Sloan *et al.* 2006), Ecology Letters (Woodcock *et al.* 2006), Philosophical Transactions of the Royal Society of London: B (Curtis *et al.* 2006), Microbial Ecology (Sloan *et al.* 2007) and FEMS Microbiology (Woodcock *et al.* 2007 In Press).

2 Literature Review

The major goal of this thesis is to develop models in microbial ecology which can be used to inform environmental engineers for the design and function of waste treatment systems. This will only be achieved through collaborative, multi-disciplinary research involving environmental engineers, mathematicians and microbial ecologists.

Accordingly, this literature review consists of three main sections. The first of these, section 2.1, provides a brief overview of the prevailing theory employed in bioreactor design for wastewater treatment. The aim is not to give an in-depth analysis of any particular process or design but rather to set the context for the research in mathematical microbial ecology. The first half of this section is devoted to the designs of systems and the second half to much of the underlying theory behind these practices. The majority of the techniques exploited take a systems perspective where, by necessity, many of the biological processes are perceived as a 'black box' and are described by empirical rules. Although these have been employed with success for many years, there is still great desire to move away from these empirical strategies and to open the biological 'black box' and describe quantitative rules for the underlying biological processes (Curtis *et al.* 2003). Without making this leap, changes in the efficiency, sizes or ability to treat recalcitrant pollutants of waste treatment systems are only ever gradual.

The current theoretical basis for bioreactor design is grounded in environmental chemistry, where the microbial communities are perceived as omnipresent catalysts. However, in all bioreactors, the self assembly of these microbial communities are replied upon, as well as their continued stability and ability to react to change. There are, however, no established quantifiable theories for these critical processes. This is despite the fact that all the system designs outlined here are utterly reliant upon microbial communities forming and metabolising waste.

Section 2.2 discusses the microbial tools which have been developed in recent years to analyse and quantify microbial systems. There are a number of exciting new technologies which are employed in microbial diversity analysis, and additional tools are being developed all the time. Each of these possible methods for laboratory work has its own relative strengths and weaknesses. Each of these must be understood and accounted for during the development of theories to ensure that the correct tools are utilised, and that the data available are useful for informing and parameterising any such models.

Finally, section 2.3 reviews a number of concepts and theories from the ecological literature. This section is given the most space in the literature review because it forms the basis of all the subsequent research. Historically, these tend to have been formulated for datasets of macroscale organisms, such as trees, birds or plants. One of the major reasons for this was that numeration of anything much smaller was, until relatively recently, not a simple task. As outlined in the previous section, current microbial tools are allowing us ever greater glimpses into the microbial world and thus are making the modelling of microbial systems an ever more manageable prospect.

Research activity at the interface between microbial ecology and environmental engineering is growing rapidly. Rittman *et al.* (Rittmann *et al.* 2006) recently presented the consensus opinions of a group of highly respected U.S. scientists working at this interface on how to create the greatest scientific breakthroughs and benefits to society. They offered what they called a 'three-peak vista' for the future of partnerships between microbial ecologists and environmental engineers. Two of these peaks pertained to improved techniques for observing and characterising microbial communities. The third peak, which was by far the least populated, was research that was orientated towards mathematical models of microbial communities. They identified the paucity of attempts to develop theory and highlighted the potential of models in both guiding the exploration of microbial communities and ultimately in the design of new treatment processes. This is a view which has been echoed by other researchers (Curtis *et al.* 2003).

Accordingly, a whole suite of new models for predicting the assembly and dynamics of communities of microorganisms are being developed. Many of these are directly analogous to established theories in classical ecology, although this newer, emerging field of microbial ecology brings with it a whole different set of challenges. For example, sampling issues for microbial systems are a problem on a vastly greater scale than those faced in macroscale studies. Understanding and overcoming such questions are key to the adaptation of current theories to microbial communities, as well as the development of completely novel models.

2.1 Wastewater Engineering

Wastewater engineering, like almost all environmental biotechnologies, is a long established field of the applied sciences, but one which may well be on the brink of a new era of understanding. Most of the microbiological processes employed today by wastewater engineers have their roots in the early part of the last century, when a series of empirical observations laid the groundwork for many of the systems still employed today (Ardern & Lockett 1914; Metcalf & Eddy 1914). Such techniques are most generally founded upon utilising established rules-of-thumb and replicating systems which have already been seen to be suitably functional.

This is not to suggest that the field has evolved devoid of any real theoretical work. However, the theoretical basis for biological systems that has been developed is anchored in the field of chemistry, rather than population biology. This approach neglects examination or the assembly and structure of the actual microbial communities which metabolise the pollutants in the wastewater, and instead examined the question of what chemical conditions within the system encourage the wastewater plants to function. Typically, these established theories relate to the energy, nutrient and environmental requirements which are observed to encourage microbial systems to flourish (Lawrence & McCarty 1971). While these philosophies have allowed for the construction of countless wastewater treatment systems, they often cannot provide any explanations of why occasional plant failure occurs. Even if chemically conditions are kept ideal for microbial communities to flourish, they cannot always predict which groups of microorganisms will establish themselves in the system.

However, the ongoing evolution and development of laboratory tools for observing and characterising microbial communities are opening up the opportunities to develop applicable theories which incorporate the microbial structures themselves. This is not to denigrate the achievements of conventional theory which have provided us with clean drinking water for decades, simply to add that a greater understanding of the microbes and their function can only improve the design and function of any future systems.

This section provides an overview and brief summary of many of the prevailing ideas and designs employed in wastewater treatment and environmental engineering.

2.1.1 Mass-Balance Equations

Traditionally, the central concept employed in the engineering design of biological wastewater treatment systems is that of mass balance. For a known quantity of waste, the nutrient and oxygen requirements of microbial communities can be estimated in order for them to successfully metabolise the pollutants (Porges *et al.* 1956). Years of empirical evidence has provided engineers with a bank of knowledge of the masses required of certain chemicals to fuel the microbiological removal of the waste.

Fundamental to the mass balance equations is the fact that during the metabolism, the total mass within the system must remain constant. That is, the sum of the masses of microbes, of nutrients and oxygen, and of pollutants during the reaction does not change. Accordingly, the net rate at which a particular reactant accumulates within any wastewater treatment system must be equal to the rate at which it is introduced to the system, plus the rate at which it is generated within the community minus the rate at which any flows out from the system (Speece & McCarty 1964).

In words, this is generally stated as the simple formula:

$$Accumulation = inflow - outflow + generation \tag{2.1}$$

It is notable that the mass balance equations, like almost all of the commonly applied theories for biological wastewater do not explicitly consider the make-up of the microbial communities relied upon. By considering all the biomass that performs a particular function as a single catalyst, without modelling the range of diversity within the microorganisms, a great deal of possibly vital information is lacking from such approaches. It goes without saying that if a particular functional group is completely absent from a system, no amount of additional nutrients will increase its growth rate within that environment.

2.1.2 Monod Kinetics

For the majority of applied microbial processes, the mass-balance equation is crucial to system design. To maximise the quantity of active biomass within the system, its growth rate should be optimised as far as possible. In practice, it is seen that for the vast majority of cases, this growth rate is limited by the substrate which acts as electron donor to the biomass.

The most commonly applied model of bacterial growth kinetics was introduced in the 1940s by the French microbiologist Jacques Monod. He noted that the growth was initially linear as a function of the limiting substrate concentration, but eventually levelled off at a constant value. Accordingly, a smooth mathematical function was proposed

$$\mu_{syn} = \hat{\mu}\frac{S}{K_S + S} \qquad (2.2)$$

where μ_{syn} is the growth rate due to synthesis, $\hat{\mu}$ is the maximal growth rate, S is the substrate concentration and K_S is a constant which gives the concentration at which the growth rate is one half of its optimal value. Implicit in this theory is the assumption that the biomass does not die or become inactive at any time when there are available nutrients.

A great wealth of empirical evidence regarding the nutrient requirements, growth rates, and rates of metabolism of microorganisms has been gathered which has allowed technologies to emerge which are generally reliable. For example, the $\hat{\mu}$ and K_S parameters in equation can be empirically observed from monocultures in the laboratory (Robinson & Tiedje 1983).

Monod kinetics have been combined with the simple mass-balance principle (Downing *et al.* 1964; Lawrence & McCarty 1971) for a theory to determine the biomass of functional groups of organisms on the basis of the available mass of chemical electron donors and acceptors. This theory forms the basis of the design of almost all environmental biotechnologies. It should be noted that, although microbial systems are implicitly relied upon for these technologies, there is no explicit description of the biodiversity within the systems. It is simply assumed that all required functional groups will be present whenever suitable resources are in the system. Understandably, this approach can prove problematic when system function declines because of biodiversity collapse, as it is completely unable to explain such behaviour.

The simplest and most commonly applied form of this theory is for a completely mixed reactor system with continuous influent and effluent (Zhou *et al.* 2002).

Applying the Monod equation (2.2) with constant outflow C per unit time gives the following equation for the rate of change of biomass, X, within the system.

$$\frac{dX}{dt} = \hat{\mu}\frac{S}{K_S + S}X - CX. \qquad (2.3)$$

This is coupled with an equation for the nutrient, S , concentration within the system where ΔS is the difference between the nutrient concentration in the influent and the effluent.

$$\frac{dS}{dt} = C(\Delta S) - \frac{\hat{\mu}}{Y}\frac{S}{K_S + S}X \qquad (2.4)$$

How Lawrence & McCarty's theory is applied depends upon the physical and operational design of reactors. Bioreactors are built according to many different designs from plug-flow, where pollutants flow continuously through the system (Schmidt 1998), to membrane reactors, which rely upon a membrane (usually now a plastic medium) which allows for the passage of treated liquids through, but which retains the vital communities of microorganisms for metabolising pollutants (Hai *et al.* 2006). The most common design however is the activated sludge process. It is not the intention here to review reactor and process designs. However, to set the theoretical research that forms the majority of this thesis into context, there is benefit in reviewing the development of the most ubiquitous technology, the activated sludge process.

2.1.3 The Activated Sludge Process

The most commonly employed biological process for wastewater treatment is the activated sludge process. Plants utilising this technique can be found in almost all climates from the tropics to the arctic regions and at a huge range of altitudes.

Its origins lie almost a century ago, when Ardern and Lockett (Ardern & Lockett 1914) noted that flocculent suspended particles formed when sewage was aerated. Such particles were referred to as being "activated" as, when retained in the system, the time for contaminants to be removed from the water was noted to be drastically reduced. Accordingly, many treatment systems were then constructed to exploit this observation (Sawyer 1965).

What is remarkable is that it was almost two decades later before it was definitively determined that the suspended flocs were indeed microorganisms and that the whole process was a biological, and not chemical one. Nonetheless, by this stage, the technique was still widely employed thanks to a wealth of empirical knowledge and parameters derived from trial and error (Sawyer 1965).

Nowadays, there are a few variant designs for wastewater treatment systems which employ the activated sludge process, but all possess the same few basic features. The first of these is the aeration tank, in which the activated sludge is mixed and kept in suspension while conditions are maintained suitably aerated so as to encourage the growth of the microbial communities within. This sludge is then passed onto the second stage of the system, the settling tank (Spellman 1997). Here, the treated water is separated from the flocs and may be returned to the environment. The flocs which are captured in the settling tank may either be wasted from the system or recycled back to the aeration tank. By this cycle of allowing the flocs to settle and then be reintroduced into the aerator, the sludge can become extremely concentrated. The wasting of some of this is done to ensure that the average sludge age, or retention time, is kept controlled.

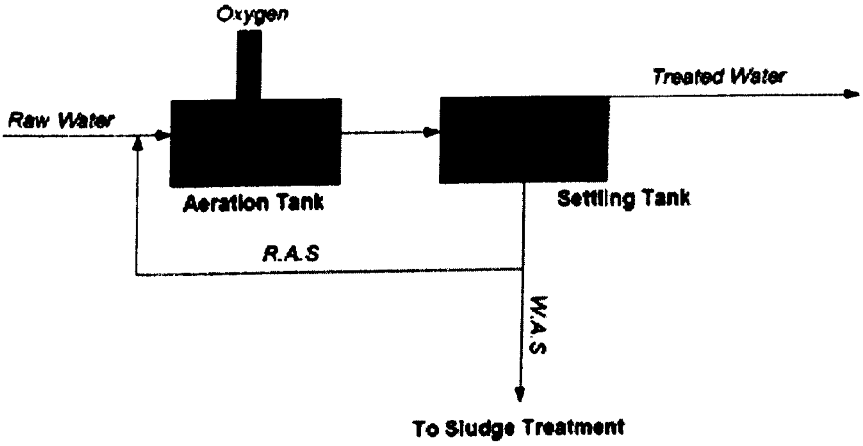


Figure 2.1

Illustration of the basic design for activated sludge plants showing the two tank system, along with the process of returning some sludge - Returned Activated Sludge (R.A.S.) and wasting part of it - Waste Activated Sludge (W.A.S.)

http://en.wikipedia.org/wiki/Activated_sludge

2.2 Molecular Microbial Ecology

Microbiology is the study of organisms too small to be visible to the naked human eye. There are two main groups of these microbes, prokaryotes and eukaryotes. Prokaryotes are characterised by the absence of a cell nucleus, and include bacteria and some types of algae. Eukaryotes possess a nucleus, and examples of microbial eukaryotes include fungi and animal-like protozoa.

Both these types of microbes are relied upon for countless everyday functions. With applications in medicine and agriculture as well as in every day food technologies, such as brewing and dairy pasteurisation, the field of microbiology is one of the most important to mankind's existence. On an even more basic level, microbial communities within the human body form the foundations of both the digestive and the immune systems.

Despite the field of microbiology being over three centuries old, it is in many ways in its relative infancy. Over the past two or three decades, the abilities of microbiologists and microbial ecologists to observe and characterise communities have been dramatically increased. Before this time, studies of microorganisms were reliant upon laboratory cultures. For such studies, microbial samples were spread across the surface of a Petri dish of nutrients and left in a laboratory to grow to higher concentrations. However, owing to the fact that some microbes grow better under such artificial laboratory circumstances than others, such culture based techniques were widely accepted as being highly likely to give biased pictures of the sample analysed. Even the most optimistic studies conceded that only a very small fraction of the microbial world could ever be cultured (von Wintzingerode *et al.* 1997); estimates of just how little vary from 0.1% to 10% of global diversity (Head *et al.* 1998). Because of this, techniques were required which did not involve the cultivation of microorganisms (Torsvik *et al.* 2002).

The other major problem faced by microbial ecologists is the lack of any consensus of a species concept for microbes (Spratt *et al.* 2006). The most commonly applied definition is based on similarity in the 16s RNA sequences, typically grouping organisms with 95%, 97% or 99% similarity as being of the same species. Clearly, two sequences which share 96% similarity could easily be classed as being of two distinct species or of being the same, depending on the similarity level chosen. Alternative definitions exist, such as defining taxonomy based on similarity in functional behaviour, such as nutrient metabolism. How employment of these differing definitions of a species concept affects the fit of certain models or theories is still being debated (Thompson *et al.* 2005).

2.2.1 Diversity Analysis Techniques

In recent years, new non-culture based tools have been developed, and they are allowing ever greater insights into the microbial world (Moffett *et al.* 2000; Daims *et al.* 2005; Kowalchuk *et al.* 2005). Thus, they are able to offer the sort of microscale data to environmental engineers that had not previously been available. This section provides a brief summary of some of the most commonly applied tools for microbial diversity analysis. There are in fact a vast number of alternative tools also in common usage, so an exhaustive review of these would prove too great an undertaking for the scope of this thesis. Those discussed here are not only some of the more common procedures, but also ones relevant to the later research presented here. A central tenet of this thesis is that in order to interpret microbial ecology surveys and translate them successfully into theory requires an appreciation of the procedures and limitations of the new molecular methods for observing microbial communities.

2.2.1.1 Polymerase Chain Reaction (PCR)

Almost all of the modern tools for analysing microbial diversity require some minimum quantity of DNA to be present in a sample in order for the techniques to function (von Wintzingerode *et al.* 1997). For microbial samples, this threshold is seldom met by the sampling procedure alone. Polymerase Chain reaction (PCR) is the most widely used method for amplifying small quantities of DNA to produce, at least in theory, many copies with exactly the same sequence (Devereux & Willis 1995).

In all PCR reactions, a few reagents are required and added before the amplification process is initiated. These include primers, which are strands of nucleic acid which anneal to the DNA, and enzymes to act as catalysts for the whole process. Once the required reagents have been added to the microbial sample, there are three stages to the procedure. The first step, known as denaturation, involves the breaking of the hydrogen bonds between the two complimentary sequence strands of the DNA in double-helix form. After this, the sample is cooled before the second stage. This is when the primers anneal to their target sequences (Sambrook *et al.* 1989; Giovannoni 1991). Finally, for the extension phase of the reaction, the sample is heated to optimise the efficiency of the polymerase, which extends those DNA fragments to which the primers have annealed.

Each time this three step cycle is followed, the DNA concentration should double. Typically samples are subjected to twenty or thirty cycles of PCR until the amplification factor has reached around 10^6 (Sambrook *et al.* 1989; Giovannoni 1991). If any higher amplification is required, the product can simply be diluted down and used as the basis for the same procedure to be repeated upon.

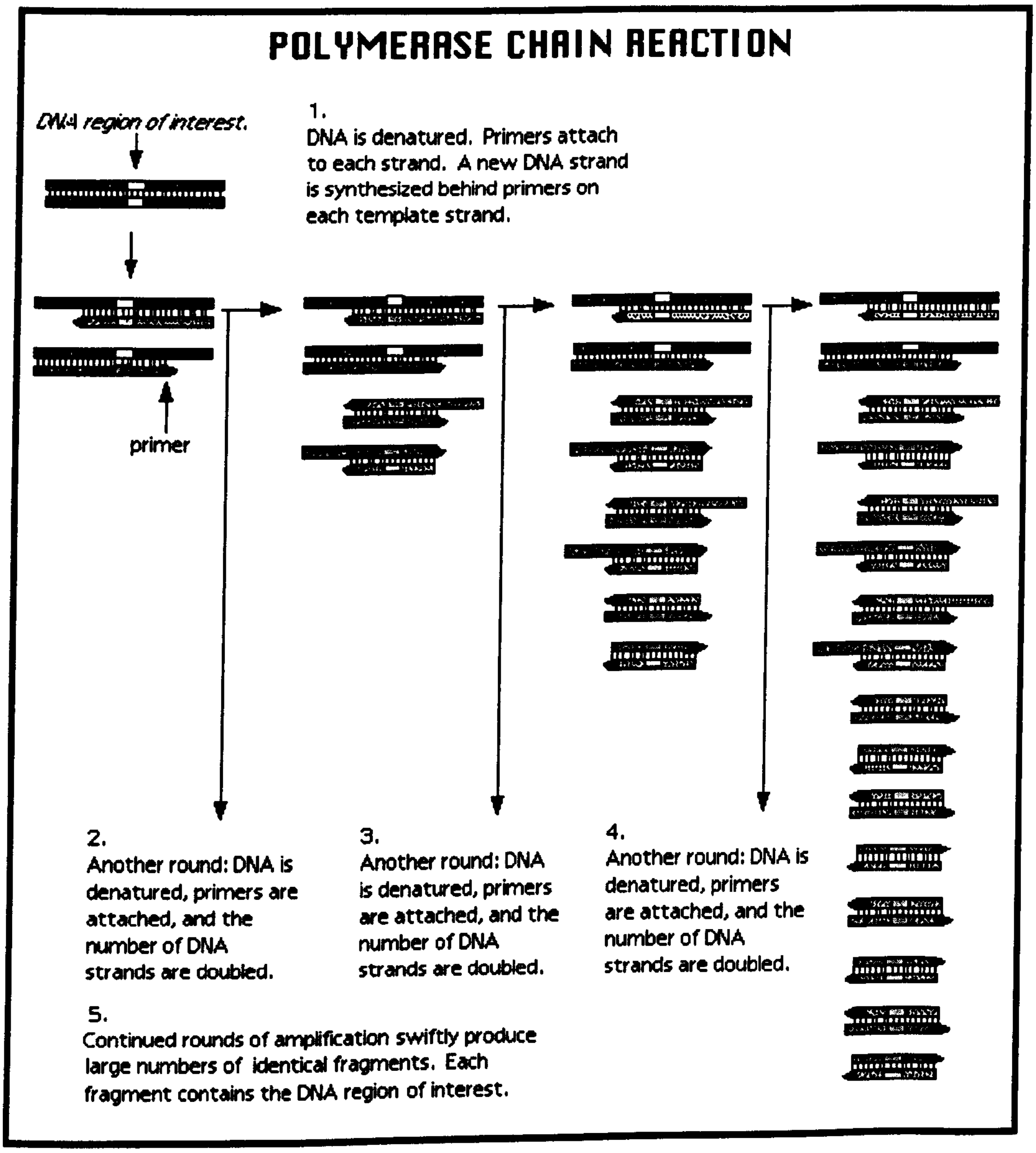

POLYMERASE CHAIN REACTION
DNA region of interest.
1.
DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.
primer
2.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.
3.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.
4.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.
5.
Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.

Figure 2.2

Diagram illustrating the procedures employed in the amplification of DNA by Polymerase Chain Reaction.

http://www.accessexcellence.org/AB/GG/polymerase.html

2.2.1.2 Denaturing Gradient Gel Electrophoresis (DGGE)

Denaturing gradient gel electrophoresis (DGGE) is a technique applied to PCR amplified DNA. After amplification, a sample of DNA is applied to an electrophoresis gel containing a denaturing agent with a gradient of concentrations, which causes the dissociation or partial melting of the double-stranded DNA fragments (von Wintzingerode *et al.* 1997; Head *et al.* 1998). The melted DNA then moves across the gel, but much slower than they would if in double helix form, and eventually come to a halt at some point.

This technique relies on the fact that different DNA fragments melt at different denaturing concentrations, so at different points across the gel. Consequently, the gel is left with many bands, each corresponding to where melted DNA fragments settle. In principle, if the DNA of each taxon melts at a different point, then each band should correspond to the presence of a different taxon in the sample (Muyzer & Smalla 1998; Muyzer 1999).

Not only does DGGE provide a measure of the number of taxa present in the original sample (by simply counting the number of distinct bands), it also can be used to quantify the abundances of each taxon. More abundant taxa in the initial sample will, perhaps somewhat unsurprisingly, result in greater concentrations of their associated DNA fragments running across the denaturing gel. Consequently, the point location at which those fragments melt will see a large build up on the gel. These show up as brightly coloured bands. A less abundant taxon will have far fewer fragments stopping at the part of the gel associated with it, and so will appear much lighter. As a result, band intensity can be measured to record the relative abundances of all detected taxa (Muyzer 1999).

Because of its simple and direct visualisation of genetic diversity, DGGE is a very common and popular tool. It is regarded as one of the most sensitive and reliable methods for community analysis (Sambrook *et al.* 1989) and one gel can analyse the structure of samples of up to around a million microorganisms, many times greater than other techniques allow. Fig 2.3 shows a typical DGGE gel arising from the analysis of 17 samples of microbes from wastewater treatment plants.

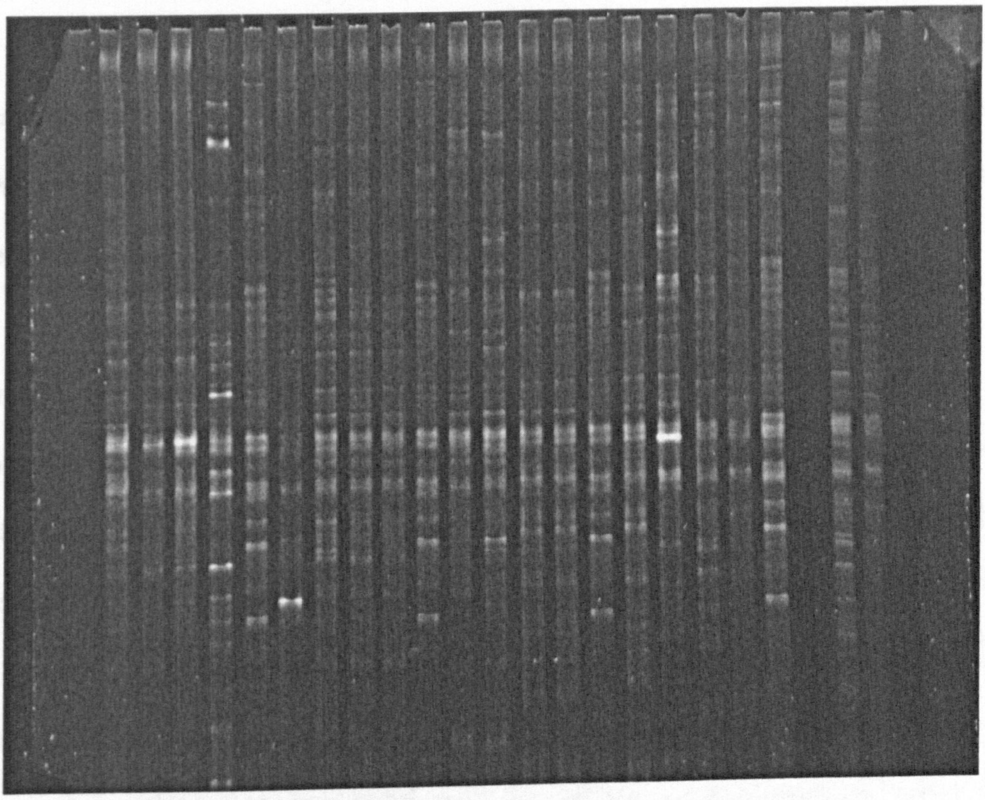


Figure 2.3
DGGE analysis of 17 samples from wastewater treatment systems. The bright bands near the centre of all the lanes clearly indicate two bacterial taxa present in high abundances in all systems. (F. Read, unpublished)

That said, DGGE is not without its drawbacks. Like any technique which has previously relied on PCR amplification, there is always the risk that biases have already been introduced (Muyzer 1999). If some sequences are better amplified than others, then the DGGE analysis will reflect this bias as much as it does the actual taxa abundances in the sample.

Additionally, there have been studies suggesting that, even aside from amplification issues, DGGE may in some cases underestimate the number of taxa present in a sample, and in others overestimate it. Infrequently, two or more

DNA fragments of differing sequences can co-migrate across the gel and stop at the same location, or indistinguishably close to each other. In rare cases, one sequence can also produce two distinct bands and hence falsely give the impression of being from the DNA of two different taxa.

The major downside of DGGE analysis, however, is its detection of rare taxa. Although it may be used to measure samples of up to around 10^6, it is very poor at detection lower ranked taxa. As identification is dependent upon a sufficient amount of DNA fragments accumulated at a point on the gel for the formed band to be seen, rare taxa may well not be able to form a band which is visible (Curtis & Craine 1998; Muyzer 1999). It is currently accepted that each taxon must make up somewhere between 0.1% and 1% of the total sample in order to be detectable on the gel (Cocolin *et al.* 2000).

2.2.1.3 Clone Libraries

An alternative to running DGGE analyses on PCR-amplified environmental samples is to construct clone libraries (Sambrook *et al.* 1989). Like other molecular methods, cloning is a multi-stage procedure.

The first stage is ligation. During this step, the DNA fragments which are to be cloned are mixed with a specially constructed plasmid, which acts as a cloning vector (Davenport *et al.* 2000). After a brief period of incubation, the PCR amplified fragments are then inserted into the cloning vector. This is known as transfection or transformation. There are various different substances which can be used as cloning vectors, but increasingly plasmids are selected which contain a so-called 'death gene' (Muneta *et al.* 1999), which ensure that unless the vector is disrupted, it will soon die.

The process of transfection is not highly efficient, and so the end result of this second stage is a mixture of the desired tranfected cells (cloning vector with inserted DNA fragment) as well as some of both the cloning vector with no inserts and DNA fragments which have not inserted into the vector. Finally the resulting cells are cultured, typically in Petri dishes. Each of the three possible substances that are cultured produce three different outcomes. The easiest to distinguish is PCR product which has not inserted into the vector. This simply will not grow and thus will not be detected in the end. The remaining cloning vector cells which have not been interrupted by the insertion of PCR products will also be killed off by their 'death gene' (Muneta *et al.* 1999), so will also be absent from the final product. Finally, where the DNA fragments have successfully inserted into the cloning vector, the 'death gene' is disrupted and the resulting product replicates in the Petri dish culture. Finally, the resulting clones can be screened and identified, usually by comparison to large publicly available database, such as GenBank. These steps are illustrated in Fig 2.4 below.

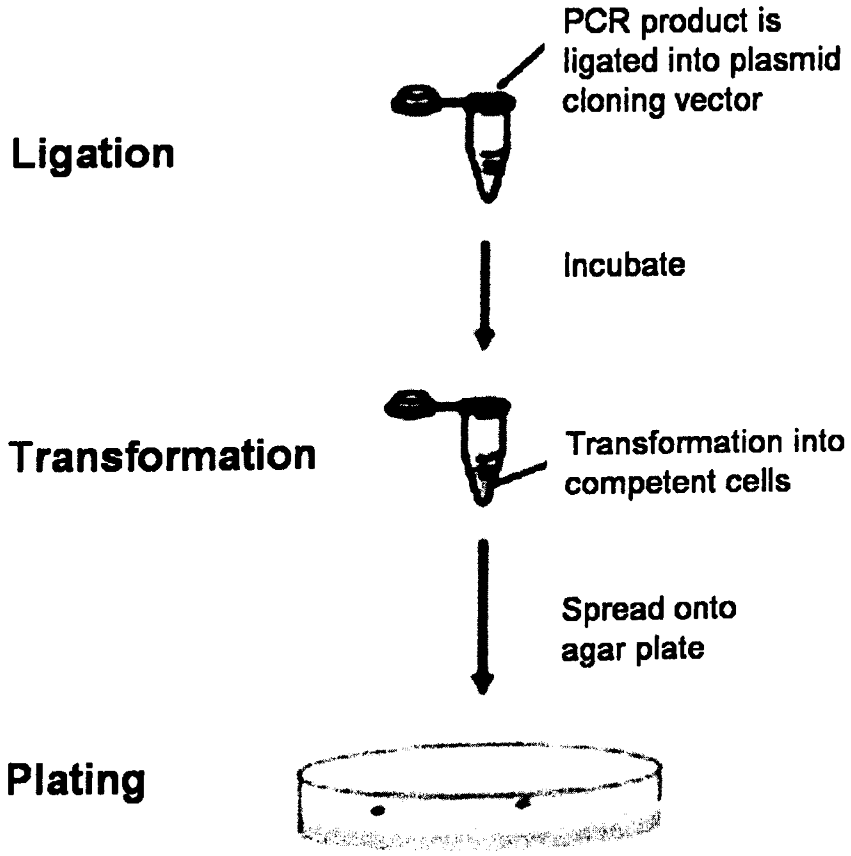

PCR product is ligated into plasmid cloning vector
Ligation
Incubate
Transformation
Transformation into competent cells
Spread onto agar plate
Plating

Figure 2.4

Schematic diagram of the steps taken in constructing clone libraries.

The major advantage of using clone libraries, rather than DGGE analysis is that there are no detection limitations. Whereas the rarer taxa in a sample analysed via DGGE may well appear absent, even a singleton in a sample will be detected in a clone library. The downsides of constructing clone libraries are the relatively high cost and low throughput. With DGGE, samples of sizes of the order of 10^6 can be examined, yet few libraries exist of even 10^3 clones. Indeed, it is not unusual to see analyses of scarcely more than 10^1 individuals with cloning.

2.2.1.4 Fluorescence *in situ* Hybridisation (FISH)

Despite the wealth of new molecular methods that have been developed, correct enumeration of microbial populations remains a challenge (Chandler *et al.* 1997). Because of the biases inherent to PCR amplification, all techniques which are reliant upon this procedure inherently cannot be accurately relied upon for quantitative studies of communities of microorganisms. There are however several PCR-independent methods for population counts, the most popular of which being fluorescence *in situ* hybridisation (FISH) (Boggs & Chinault 1997).

FISH uses fluorescently labelled probes to detect specific organisms in biological samples (Amann *et al.* 1990b). It is an especially valuable tool when interested in the dynamics of a population through both space and time.

FISH is carried out in four main steps, commonly known as fixation, hybridisation, washing and fluorescent microscopy. The first of these steps is carried out to ensure that the cells within the biological sample retain the morphology and basic structure during the subsequent procedures. Typically, this is achieved with the addition of some fixative, most usually an alcohol or aldehyde. Once fixed, the cells are exposed to fluorescently labelled probes, which then bind to the RNA of metabolically active cells (Amann *et al.* 1990a). These probes have to be carefully selected to be specific to the groups of microorganisms of interest, so that they do not bind to other non-target cells. After hybridisation, the sample is then washed to remove any excess probe which had not bound to target cells. Finally, the washed sample can then be viewed under an epifluorescence microscope or confocal scanning laser microscope. The target cells to which the probe has bound will then be excited by the laser in the microscope and will fluoresce. Numeration is then done by a simple count of the number of fluorescing cells in the resulting image.

Figure 2.5 below illustrates a typical image which is obtained via FISH.

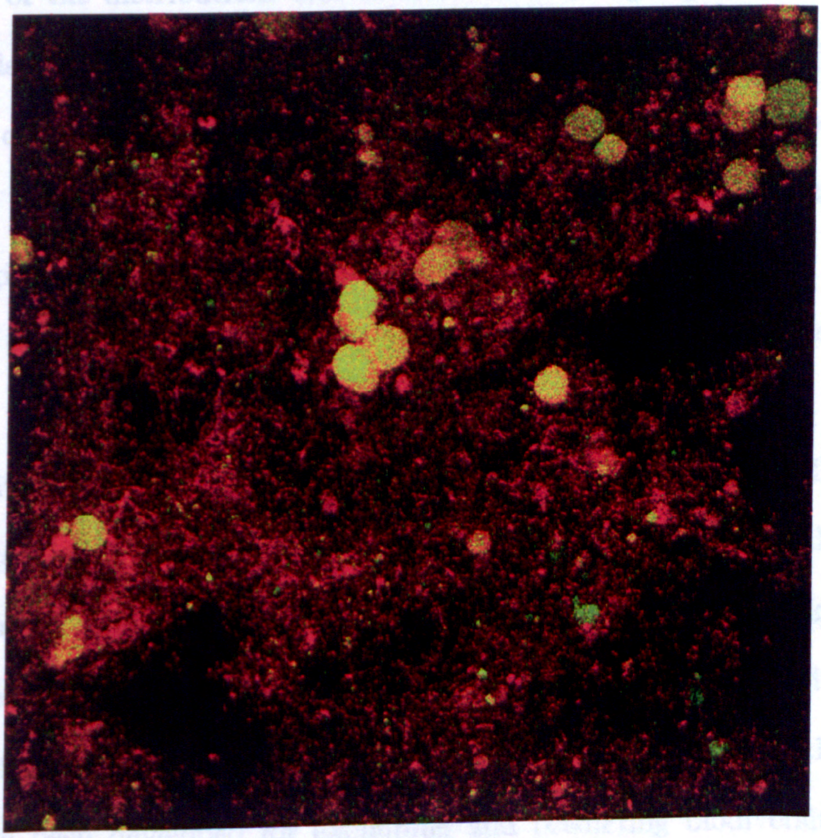


Figure 2.5

A typical image of FISH signals obtained using an epifluorescence photomicroscope. (R. Davenport, published).

The main advantage of FISH is that it is possibly the best available tool for enumeration of microbial samples (Waar *et al.* 2005). It is, however an extremely slow and therefore expensive method which somewhat limits its spread of application.

2.3 Ecology

The study of the distributions and abundances of living organisms is one of the oldest fields of the biological sciences (de Candolle 1855; Jaccard 1901). It is borne out of a natural curiosity to enquire how many different species will be found in a given habitat and at what abundances, and more generally how and why biological communities assemble. Accordingly, over the years, there have been many surveys undertaken on the distributions of birds, plants and animals.

Until relatively recently, these studies have been confined to such large multi-cellular organisms (mammals, birds fish etc), which in this thesis will be termed macro-organisms. The major reason for this was simple; observing microbial communities in their natural environment was extremely difficult and taxonomic classification even harder. Consequently, the majority of published ecological models have been developed for explaining and predicting biodiversity patterns for these larger organisms. However, as previously mentioned, a wealth of new techniques are being developed which are affording researchers the chance to study the ecology of microbial communities *in situ*. As these new tools improve and allow for the gathering of ever more extensive datasets from microbial communities, the potential for the development of a similar suite of models for microorganisms grows ever greater.

This section provides an overview of some of the main concepts and theories currently employed in classical ecology, as well as discussions of how some of these may be adapted to form the basis of predictive microbial models. The aim of this review is certainly not to provide an exhaustive summary of ecological theories, rather to cover those which may be useful in laying the groundwork for similar models for communities of microbes.

2.3.1 Key Concepts in Classical Ecology

2.3.1.1 Species-Area Curves

One of the oldest and most studied concepts in ecological literature is the relationship between the area of a habitat and the number of distinct species housed within (Arrhenius 1921; Preston 1957).The most commonly cited form for this is that of a positive power-law $S \propto A^z$relating the number of species$S$observed in a given area$A$. Studies for a wide range of organisms, from the scale of ants to trees, have found that the positive exponent, z, tends to be between 0.16 and 0.35. In other words, the number of species encountered is typically proportional to somewhere between the third and the sixth root of the area examined.

This seemingly universally applicable, simple formula has had a huge impact on ecological practice and theory. In practice, it has been touted (Diamond and May, 1981), perhaps controversially (Saunders *et al.* 1991)as a framework for assessing the long term effects of the fragmentation of ecosystems, or other reductions in habitat area on species diversity. From a theoretical perspective it spawned many developments (Sugihara 1980) including MacArthur and Wilson's Theory of island biogeography, (Macarthur 1960) (section 2.2) widely perceived as one of the cornerstones of modern ecological theory.

Species area relationships are generally illustrated by plotting the patterns on a plot using a log-log scale. This simple transformation of variables gives, for the $S \propto A^z$ power law, a straight line fit

$$\ln S = z \ln A + \ln S_0$$

so that the gradient on the log-log scale gives the value of the z exponent.

Figure 2.6 below shows a typical species area curve for a study of plant species in California.

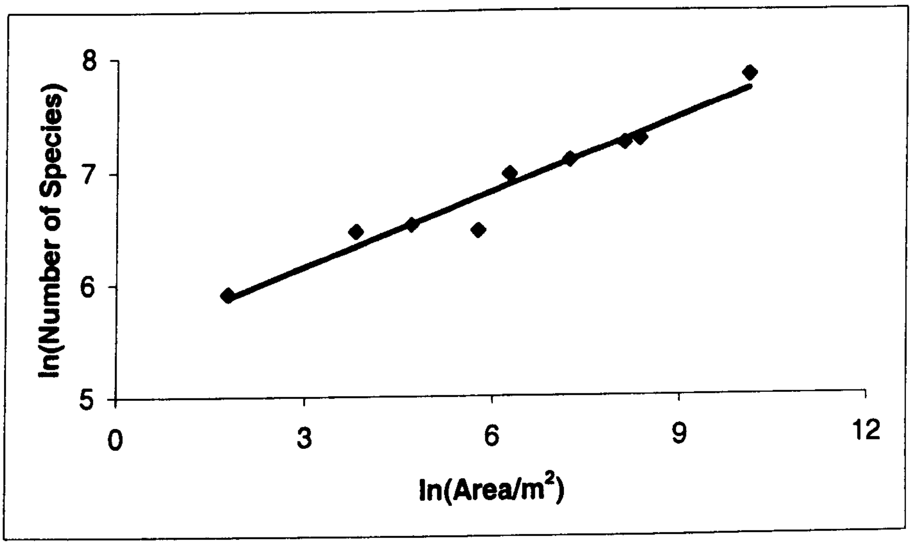


Figure 2.6

Species area relationship for endemic vascular plant species found in California. The value of the *z* exponent is 0.22.

Redrawn from Johnson et al. (1968)

In practice, there are several different types of species area relationship (Scheiner 2003) which are studied. The three most commonly investigated types involve nested sampling, grid based sampling, and island-like sampling schemes. These are illustrated in fig 2.7. Somewhat predictably, all types of species area relationship give, almost without exception, a positive correlation between the area examined and the number of species detected.

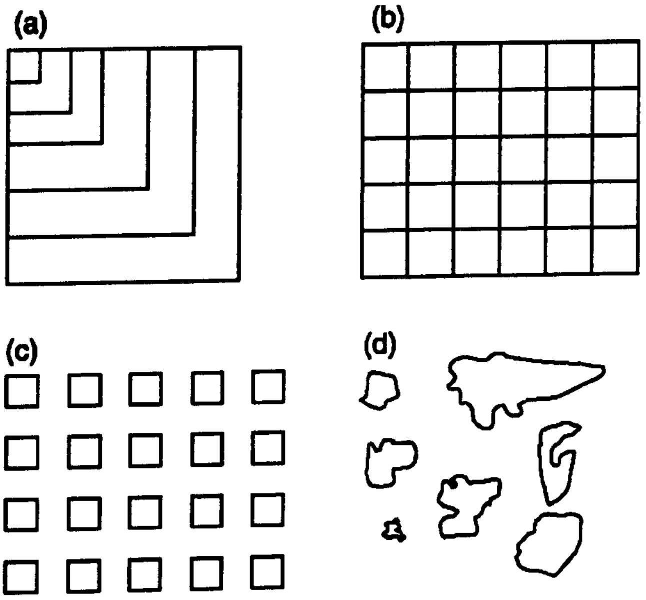


Figure 2.7
Species-area relationships can be built from four main sampling schemes (a) Nested sampling, (b and c) Grid based sampling, either contiguous or non-contiguous and (d) Island-like. From (Scheiner 2003).

However, there are two very different mechanisms which drive this observed phenomenon. The first is a purely statistical reason. For nested sampling and island-like sampling schemes, where ever larger areas are considered, it is clear that sampling a greater number of individuals will give an increased chance of encountering new additional species to the study. This is especially true in cases when the dispersal of species is somewhat limited.

The second major influence driving the observation of such relationships is a biological reason; larger regions of habitat tend to be more environmentally heterogeneous, thus different species are likely to be encountered at different locations within the region of interest. This mechanism is the underlying process behind grid based species-area relationships, where samples are taken of equal size and behind some techniques for both island-like and nested sampling schemes, which rely on point measurements and the distance-decay of similarity between samples (Harte *et al.* 1999; Green *et al.* 2004b).

Since the study of microorganisms is in its relative infancy, it is perhaps unsurprising that microbial ecologists have alighted on the species-area relationship in their search for general patterns. There have been a number of studies recently that identify taxa-area relationships for microbes in both contiguous (Green *et al.* 2004b; Hughes Martiny *et al.* 2006) and island-like communities (Bell *et al.* 2005a). From an engineering perspective these relationships, if found to be sufficiently consistent, could assist in designing for diversity based on the area and volume of bioreactors.

2.3.1.2 Species Abundance Distributions

Although studies of the species-area relationship provide information on the number of different species found at a given site, they do not address the structure of the community. Knowledge of species richness does not shed any light on the relative abundances of those species present. It is common to find two communities which house the same number of species, but whose communities are vastly different in structure (Simpson 1949; Pielou 1966).

For example, some systems are seen to consist of a number of almost equally abundant species, for example bacteria in soils (Borneman & Triplett 1997). On the other hand, communities exist which can have similar total numbers of species, but which are dominated by a few highly abundance species, with all others observed only rarely (Corbet 1941; Spratt *et al.* 2006). Systems for which all species abundances are similar are regarded as being evenly distributed, and ones dominated by a few top ranked species are uneven. A number of different indices are applied to quantify this concept, but no one measure is regarded as 'standard' (Shannon & Weaver 1949; Pielou 1966).

This concept of evenness is of paramount importance to certain fields of ecology, as well as to environmental engineers working with biological systems. It is often observed that evenness within a functional group promoted stability of that function (Rowan *et al.* 2003) or productivity (Bell *et al.* 2005b). Biotechnologies which rely upon the presence of certain bacterial taxa (or functional groups of them) require not just the maintenance of these within the engineered system, but also the preservation of them at suitably high abundances to metabolise the required waste. In the context of wastewater treatment plants, for example, an extremely uneven community structure is more likely to suffer loss of function.

The relative abundances of taxa in a biological community are generally expressed in the form of the species abundance distribution which is often illustrated in a number of different ways, most notably as a species-abundance distribution and a ranked species-abundance distribution.

A species-abundance distribution is a histogram that counts of how many species have only a single organism in the community, how many have two organisms there, and so on. In practice, many species abundance distributions seem to have strong positive skew; that is, the communities consist of a small number of very abundant species and have a large number of rarer ones (Corbet 1941), (Preston 1948). Consequently, the histograms are unclear as much of the data is closely packed at the left hand of the plot, with a huge narrow tail extending out to the right (Spratt *et al.* 2006). For such datasets, it is advantageous to employ logarithmic binning of the abundances, a technique first introduced by Preston. Preston argued that since studies are often concerned with how populations double in numbers, a natural base for logarithms would be base 2. This convention remains to this day in the ecological literature.

Once the simple histogram count has first produced based on the abundance of each species within the community, these data are then sorted into bins with edges 0,1,2,4,8,... When any particular species abundance falls exactly on the bin edge, it is divided equally between the two bins either side of the boundary. For example, the bin labelled 8 contains the number of species with 5, 6 or 7 organisms along with half the number with 4 and half the number with 8.

A ranked species abundance distribution essentially plots the same information in a different manner. It is conventionally used when comparing communities where the distribution of taxa are quite different (May 1975). The species are ranked 1 to n according to their relative abundance, where n is the total number of species observed, rank 1 is the most common species and rank n is the rarest. The relative abundances are then plotted against their given rank, producing a monotonically decreasing ranked abundance distribution.

For extremely large communities, the species abundance distribution is often approximated to a continuous function. Strictly speaking, the distribution must be discrete, as the number of species at a given abundance level must be a non-negative integer. However, when community sizes and species richness are high, small rounding errors can be ignored and it is possible to regard the species abundance distribution as being continuous. Just as for continuous random variables in probability theory (Grimmett & Welsh 1986) , where it is only possible to state a probability density function (pdf) for a distribution rather than individual discrete probabilities of each possible state, a species density function is required for the abundance distributions using a continuous approximation. The number of species at a particular abundance level is given by an integral across the species density function rather than a sum of the counts in appropriate bins, in the same way as expectations of continuous random variables are found by integrals across the required domain rather than summations over all possible outcomes.

Displaying the data as histograms or as ranked abundance distributions gives a visual picture of the distribution. Frank Preston and other eminent biologists of the time including RA Fisher (who went on to become one of the pre-eminent statisticians of the twentieth century) were immediately drawn to the idea that it might be possible to fit a parametric distribution to these data with a view to classifying communities and ecosystems (Preston 1948; Spratt *et al.* 2006). These models could be perceived as being phenomenological in that they were originally proposed without any attempt to reconcile them with any underlying biological community assembly process. Ultimately, it was thought that these might lead to an understanding of how the communities formed and functioned and biological explanations have proffered post-hoc.

Many people, however, frowned upon such a "statistical" approach to finding the distribution of individuals per species within a community. One of the most prominent of these critics was Robert MacArthur, who argued that is the same mathematical functions were so ubiquitous in describing real ecological systems, there must be some underlying mechanisms of community assembly and dynamics upon which they could shed light (Macarthur 1960). In the wake of such publications, a second class of models followed that were based on conceptual models of how communities formed. To this day, there is still a great deal of debate over the relative merits and applicability of these two classes of models (Harte 2003; McGill 2003).

2.3.2 Phenomenological models of Community Composition

Through the history of ecology there have been many different mathematical functions proposed as models for observed taxa abundance distributions and it is not feasible to review them all. However, there does appear to be a consensus, although not one uniformly held, that most data can be described by one of three common distributions; the lognormal, the logseries and the geometric.

The majority of the theory developed for explaining ecological communities is centred upon explaining the species abundance distributions which are observed in real datasets. As a consequence, there are two very different classes of models which have arisen to explain ecological systems. The first of these has its origins in developing a distribution which explains the species abundances directly. Such models rely on finding mathematical functions which can replicate the observed shapes of species abundance distributions. For many of these models, subsequent ecological explanations have been proffered as to why such patterns should arise, but such justifications are often dismissed as being contrived or having only weak biological foundations (Boswell & Patil 1971; Pielou 1975).

2.3.2.1 The Logseries Distibution

One of the earliest suggested mathematical functions fitted to observed species abundance data was that of a logseries distribution. It was observed (Spratt *et al.* 2006) that the distribution of abundances for a dataset of 620 species of butterfly seemed to form a smooth hyperbolic progression. Accordingly, a negative binomial distribution was fitted to the dataset. However, owing to the fact that the zero abundance class was unobservable, the distribution was truncated to remove this category. Under the additional assumption that the number of species present could be effectively infinite, a one parameter distribution was suggested; a truncated negative binomial distribution as the shape parameter tended to zero.

The result was the logseries distribution which states that the number of species expected with n individuals in the community is

$$\theta\frac{x^n}{n}. \tag{2.5}$$

The original notation used by Fisher uses the scaling parameter α rather than θ. However, as discussed later in this chapter, the logseries is of interest to more recently developed neutral community models and in that context, the alternative θ notation is the convention.

Although this may seem at first glance to be a two parameter model (both $\theta\,\text{and}\,x$) one of the parameters is constrained by the other and by the total number of individuals, N_T, in the system.

Equating the total number of individuals

$$N_T = \sum_{n=1}^{\infty} n\left[\theta\frac{x^n}{n}\right] = \theta\frac{x}{1-x} \tag{2.6}$$

thus

$$x = \frac{N_T}{N_T + \theta}. \tag{2.7}$$

When plotted without using Preston's logarithmic binning, the logseries distribution resembles a familiar hyperbolic shape, as shown in figure 2.8 below. When plotted on a logarithmic scale, the shape is much flatter, often almost constant across the first few bins before tailing off rapidly when the bins for the small number of highly abundant organisms are reached (figure 2.9).

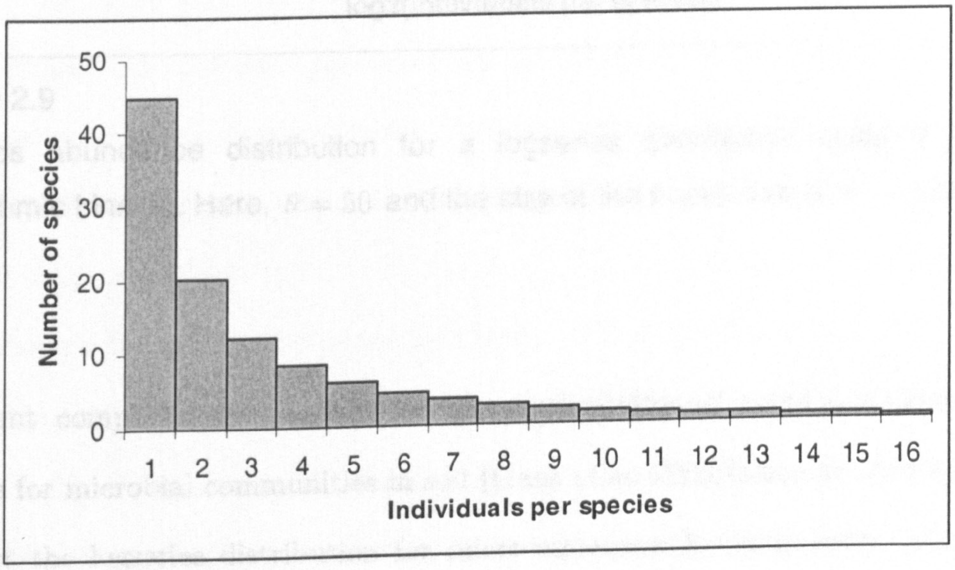


Figure 2.8
Species abundance distribution for a logseries distribution where $\theta = 50$. The size of the population is $N_T = 450$.

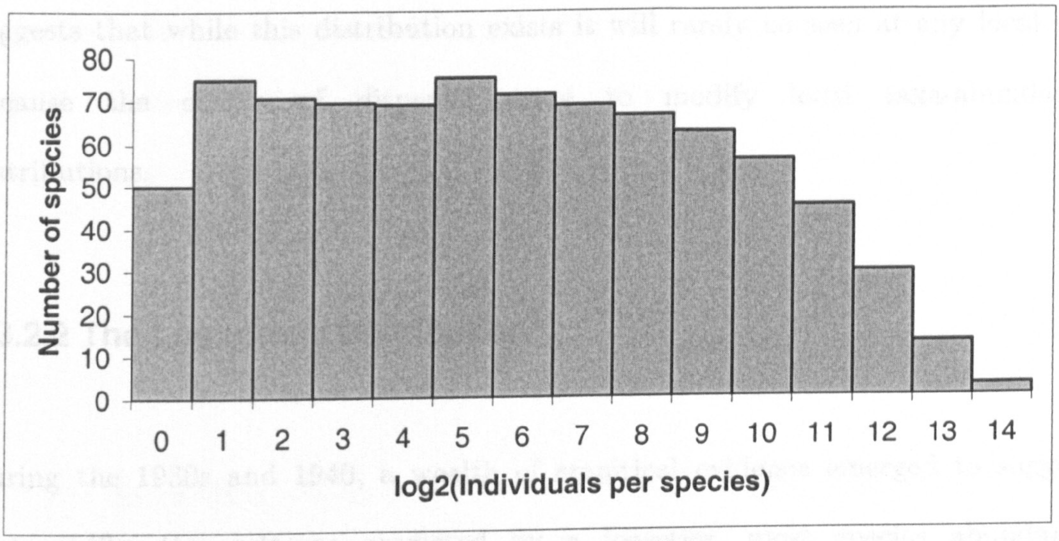


Figure 2.9
Species abundance distribution for a logseries distribution using Preston's logarithmic binning. Here, $\theta = 50$ and the size of the population is $N_T = 17000$.

A recent comprehensive analysis of the applicability of various mathematical models for microbial communities in soil (Gans *et al.* 2005)found little evidence to support the logseries distribution for micro-organisms living in soils. Interest in the logseries distribution was short lived and few, even heuristic, justifications for the logseries distribution were ever advocated (May 1975). However, it is included here because there has been a renewed interest brought about by neutral theories of community assembly which are discussed in detail later in the literature review.

In Hubbell's version of the neutral theory (Hubbell 2001) he advocates a conceptual model of speciation and extinction that leads to a logseries distribution of taxa abundances for large biological communities. He further suggests that while this distribution exists it will rarely be seen at any local site because the effects of dispersal serve to modify local taxa-abundance distributions.

2.3.2.2 The Lognormal Distribution

During the 1930s and 1940, a wealth of empirical evidence emerged to suggest that, unlike the patterns predicted by a logseries, most species abundance distributions had an internal mode (Williams 1964). This led to the logseries falling out of favour with many theoretical ecologists.

Furthermore, it was initially noted by Preston (Preston 1948) that when datasets were plotted using logarithmic binning the histogram resembled the classic normal bell-shaped curve. He therefore proposed that species-abundances were distributed lognormally. Since then it has become the most commonly cited distribution for species abundances. Lognormal distributions have been fitted to species-abundance data in plants (Pielou 1975), mammals (Preston 1962), fish (Magurran 1996) and birds (Price *et al.* 1995; McGill 2003).

Some authors (McGill 2003) have suggested that it is so ubiquitous it should be used as a null hypothesis on the distribution of species abundances against which any other proposed model is compared .

Mathematically, the lognormal is a continuous distribution with probability density function (pdf) such that if $X \sim LogN(\mu,\sigma^2)$, where μ and σ are the mean and standard deviation of $\ln(X)$ then

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left(\frac{1}{2\sigma^2}(\ln(x)-\mu)^2\right). \tag{2.8}$$

The distribution is so named because if $Z = \ln(X)$ and $X \sim LogN(\mu,\sigma^2)$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{1}{2\sigma^2}(z-\mu)^2\right). \tag{2.9}$$

That is, Z has a normal distribution $Z \sim N(\mu,\sigma^2)$.

Because of the convention in ecological literature introduced by Preston, who was the first to advocate lognormal distributions, of organising data into logarithmic bins using base 2, the lognormal species abundance distribution is normally stated in a slightly different form. Let S_0 be the number of species in the modal bin, S_1 the number in the bin to the right of the mode, S_{-1} the number in that one to the left of the modal bin and so on. It is then observed that these histogram counts are distributed according to

$$S_R = S_0 \exp(-a^2R^2) \tag{2.10}$$

where a is an inverse measure of the variance of the distribution.

While it may not be instantly apparent that this distribution is of the form of (2.8), setting the simple change of variables

$$R = \log_2 N - \log_2 N_0$$

and

$$a = \sqrt{\frac{\ln(2)^2}{2\sigma^2}}$$

gives a relationship of the form (2.8), up to a multiplication factor of $S_T = \frac{S_0 a}{\sqrt{\pi}}$ and thus is indeed lognormal. The multiplier is there so that the complete integral indeed gives the total number of species, S_T rather than simply unity.

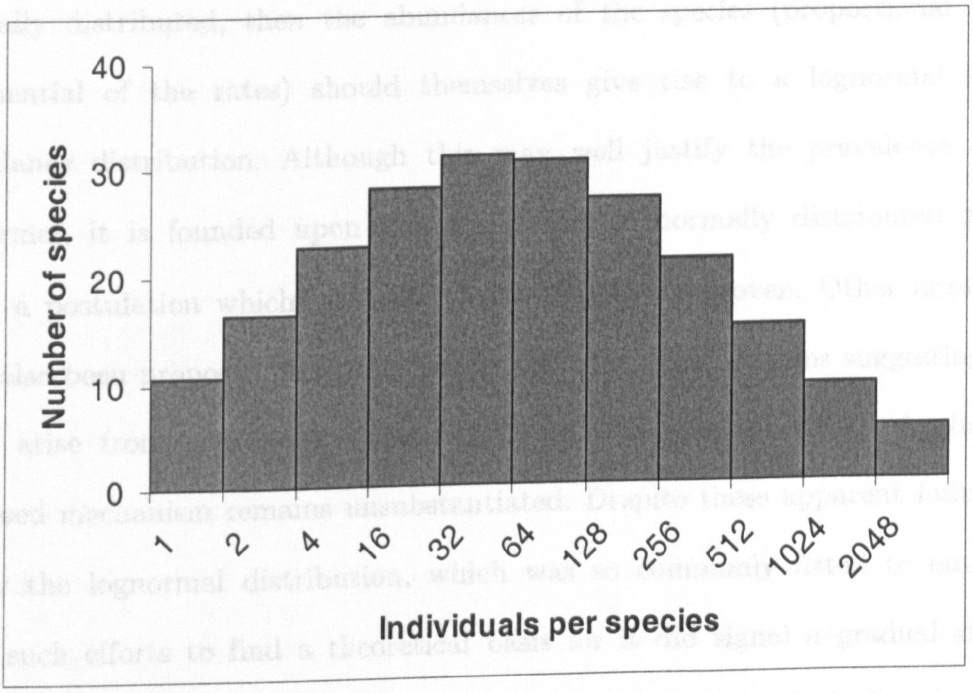


Figure 2.10

Species abundance distribution for a lognormal distribution where $S_T = 225$, $N_0 = 20.82$ and $\sigma = 2.98$. These parameters are the best fit calibrated for a community of trees on Barro Colorado Island, Panama (Volkov *et al.* 2003).

The ubiquity of the lognormal distribution led some eminent ecologists to believe that there must be some underlying process that led to the distribution(May 1975; Caswell 1976). As a consequence they sought to derive a set of rules for community assembly that would yield a lognormal distribution in taxa abundances for a community.

The proposed explanation for the lognormal which has been most commonly cited is that the distribution arises from a number of random processes affecting the growth of populations. It is argued (May 1975) that if these growth rates are normally distributed, then the abundances of the species (proportional to the exponential of the rates) should themselves give rise to a lognormal species abundance distribution. Although this may well justify the prevalence of the lognormal, it is founded upon the assumption of normally distributed growth rates, a postulation which has never been definitively proven. Other arguments have also been proposed, including purely statistical explanations suggesting they could arise from combining unrelated samples (Routledge 1980). Again, this proposed mechanism remains unsubstantiated. Despite these apparent failures to justify the lognormal distribution, which was so commonly fitted to empirical data, such efforts to find a theoretical basis for it did signal a gradual shift in ecology. Rather than simply considering which mathematical function best described existing datasets, the underlying processes that led to species abundance distributions were being considered. From the point of view of those seeking to predict or manipulate and engineer biological systems, this was a highly important and substantial change in philosophy.

There has, however, as yet been little conclusive evidence to support any particular distribution for microbes; although many studies (Gans *et al.* 2005) (Dunbar *et al.* 2002) are slowly whittling away at the set of feasible distributions. Furthermore, some authors have reasoned that there will be a characteristic species-abundance distribution for a particular environment and then proceeded to speculate on the consequences for the diversity of ecosystems under particular assumptions on the form of the distribution. One of the most highly cited of these is the $N_T / N_{\max}$ ratio method developed by Curtis *et al.* (Curtis *et al.* 2002). They derived a method for crudely parameterising the distribution based on data that microbial ecologists could easily measure. Given only values of the number of individuals in a community, N_T, and the abundances of the most and least common taxa ($N_{\max}$ and $N_{\min}$ respectively) they developed a quick and simple rule of thumb for estimating diversity. Under the additional conservative assumption that $N_{\min} \approx 1$, the technique can provide diversity estimates based solely on the total population size and the ratio $N_T / N_{\max}$. Based on, this they were able to estimate the total diversity in samples from many environments and concluded that some environments such as soil could be astronomically diverse (10^5 per mg). The underlying assumptions of the Curtis *et al.* approach have not been confirmed. However, the study did serve to highlight the utility of taxa-abundance distributions in extrapolating from small samples of micro-organisms to estimate community richness.

2.3.3 Common Models of the Assembly of Ecological Systems.

In addition to some of the phenomenological models of community structure, there is also a category of theories which seek to explain the underlying biological mechanisms behind community structure. From the point of view of those seeking to engineer biological systems, this second category is almost certainly of greater use. If explanations can be offered concerning the way in which such communities assemble and change through time, then perhaps these models can also help suggest how systems may be engineered to produce communities with certain desired properties. The species abundance distributions themselves are often implicit in such models, along with information on other attributes of the communities, such as the dynamics or the stability of any patterns.

2.3.3.1 Theory of Island Biogeography

Whilst many ecologists were following MacArthurs's lead and pursuing theoretical explanations for taxa-abundance distributions (Kilburn 1966), MacArthur and Wilson took a step back and decided to explore theoretical explanations for the taxa-area relationships. They took what at the time was a radical perspective in that they considered dynamic processes of colonisation and extinction on Island communities rather than attributing all of community composition to the availability of resources or local growth alone. What emerged was the Theory of Island Biogeography, first published in 1967 (Macarthur 1967). This remains probably the most enduring and commonly cited explanation of the factors which govern the assembly of multi-species communities on islands. In the context of their theory, 'islands' need not simply be areas of land surrounded by sea, rather any such habitat with clearly defined boundaries. Accordingly, the theory could be applied to plant species in an enclosed woodland area (Lavin *et al.* 2001) or to fish species within a given pond (Browne 1981; Tonn & Magnuson 1982).

The conceptual picture which underlies the theory is one of biodiversity within an island being determined by a series of randomly occurring events; the chance extinction of species currently on the island and the possibility of new species migrating in from the mainland or from a neighbouring island. In its most basic form, the theory posits that, in the equilibrium state, the number of different species housed on an island is governed by the balance between two opposing factors; isolation and island/habitat size.

The isolation of an island, or how distant it is from others, determines the rate of immigration from other similar communities onto the island. For islands that are very isolated, with few similar islands in their surrounding area or that are distant from the mainland, the immigration of new species into the habitat is assumed to be low. When there are many such neighbouring islands, individuals can more readily pass between close-by habitats, and thus the rate of immigration into each island is much higher. With higher immigration of individuals comes the increased probability that new species will be introduced into a community by migration to the island, and thus that the diversity within will be increased.

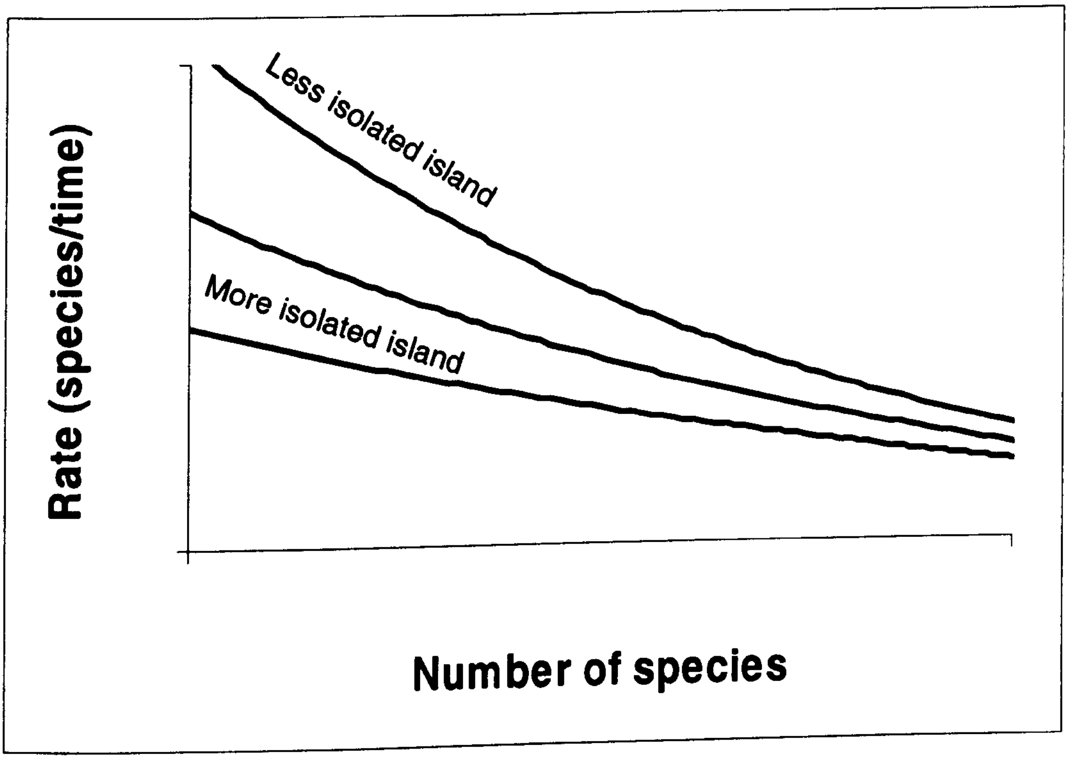


Figure 2.11
The immigration rate of new species onto an island according to the Theory of Island biogeography for islands of three different degrees of isolation. The more isolated the island, the lower the rate of new species immigrating.

The second key component in determining the biodiversity which can be maintained on a given island is the island's size. Just as an 'island' may be defined to be a pond or a woodland area rather than a landmass surrounded by water, the 'size' of an island need not be solely a measurement of physical space, such as volume or area. In most contexts, larger islands are seen to house more individuals because of both increased physical space and a larger pool of resources. However, in some cases this may not hold. For example, a habitat rich in nutrients would be considered a 'larger' site than one of equal physical proportions but far more scarce in resources. In other words, the 'size' of an island is really a measure of how many organisms it can sustain. This factor then has a knock-on effect on the rate at which species become locally extinct. Larger islands, which house more individuals, are less likely to see species disappearing completely from the habitat. The extinction rate is therefore lower for islands of greater size.

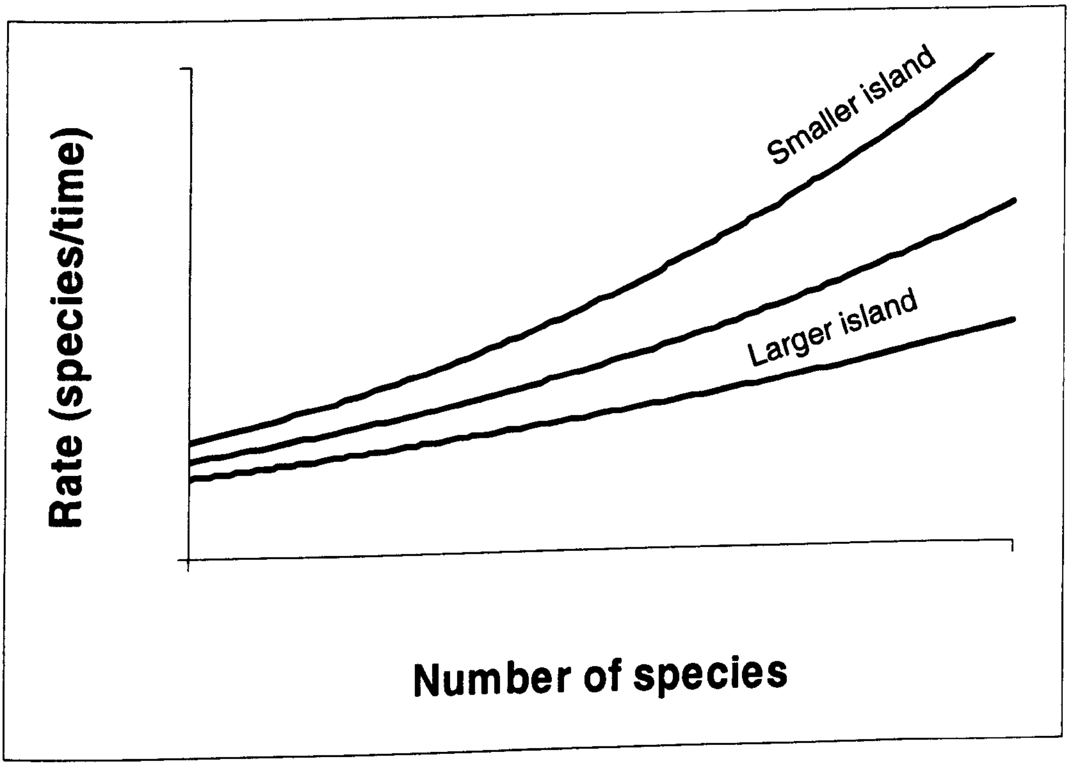


Figure 2.12
The extinction rate of current species on an island according to the Theory of Island biogeography for islands of three different sizes The larger the island, the lower the rate of local species loss.

According to the Theory of Island Biogeography, the steady-state biodiversity on a given island is determined by the balance between these two opposing influences; extinction and immigration. By definition, an island is in its equilibrium state of biodiversity when the net change in species richness is zero. This occurs when the extinction rate of species from the island is equal to the rate of immigration of new species into the area. Graphically, if both the extinction rate (species loss per unit time) and the immigration rate (new species arriving via immigration per unit time) are functions of the population size, then the steady-state biodiversity will be determined by the point at which these two curves intersect.

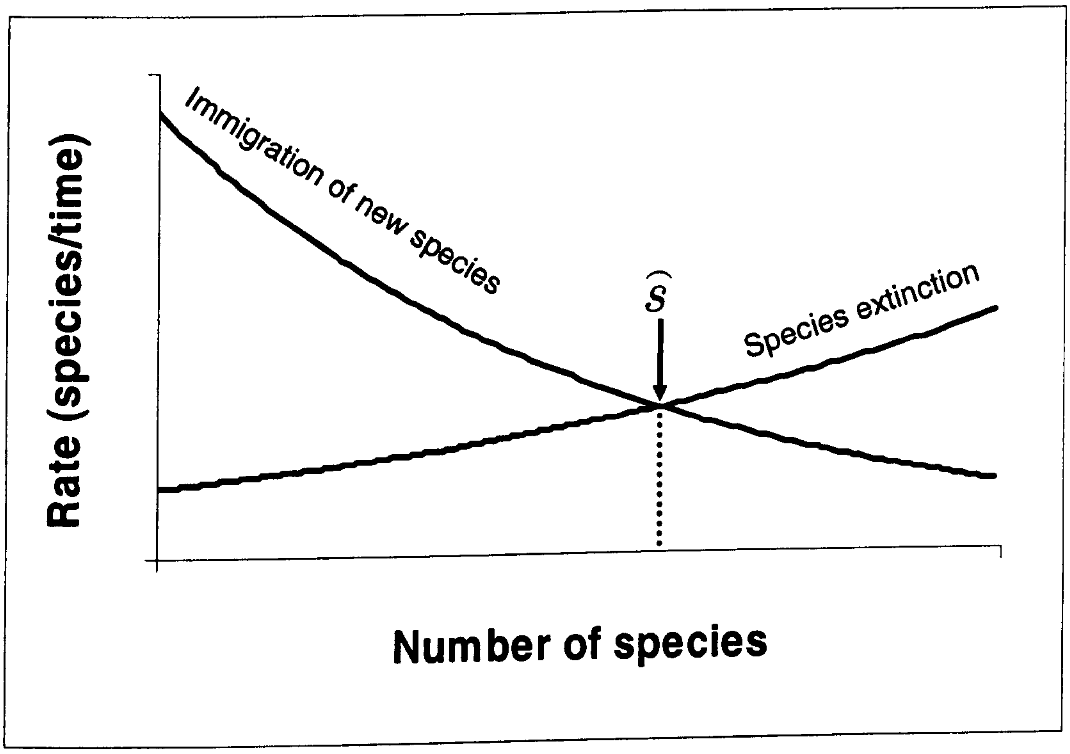


Figure 2.13
Equilibrium model for species richness as predicted by the Theory of Island Biogeography. The predicted number of species, $\hat{s}$, is given by the intersection of the immigration and extinction curves.

At a first glance, the Theory of Island Biogeography seems appealing to microbial ecologists. Communities of micro-organisms are most usually open and subject to immigration from air- or water- borne individuals which may establish themselves within the system. It, therefore, might seem that the theory would be ideally suited to microbes. However, MacArthur and Wilson's theory produces predictions of the total global diversity. When microbiologists cannot even agree on the order of magnitude of the diversity within even a single gram of soil (Volkov *et al.* 2003; Gans *et al.* 2005; Bunge *et al.* 2006) predictions on such a scale will prove to be utterly unverifiable. A slightly different theory would be required for microbial systems, one which is capable of predicting diversity at a variety of scales. Bell *et al.* (Bell *et al.* 2005a) recently observed more diversity with increasing community size in insular treehole communities. However, their diversity estimates were determined using DGGE, with which only a small fraction of the total richness can be seen (section 2.1). They did not attempt to explain their observations in terms of Island Biogeography, rather they speculated on the distribution of niches in the various 'islands'.

Although the applicability of such theories to microbial systems is widely accepted (Hughes Martiny *et al.* 2006) there are still some who believe that there is no real biogeography in the microbial world (Finlay & Clarke 1999). They argue that dispersal of microbes is so widespread and community sizes sufficiently large as to preclude any local stochastic extinctions. This "everything is everywhere" hypothesis is certainly controversial and the majority of published studies do indeed detect some microbial biogeography (Green *et al.* 2004a; Bell *et al.* 2005b; Hughes Martiny *et al.* 2006). The work presented later in this thesis (chapter 5) also finds evidence to support the idea that the dispersal of microbes is not unlimited and that migration events may well be key to shaping microbial community structure.

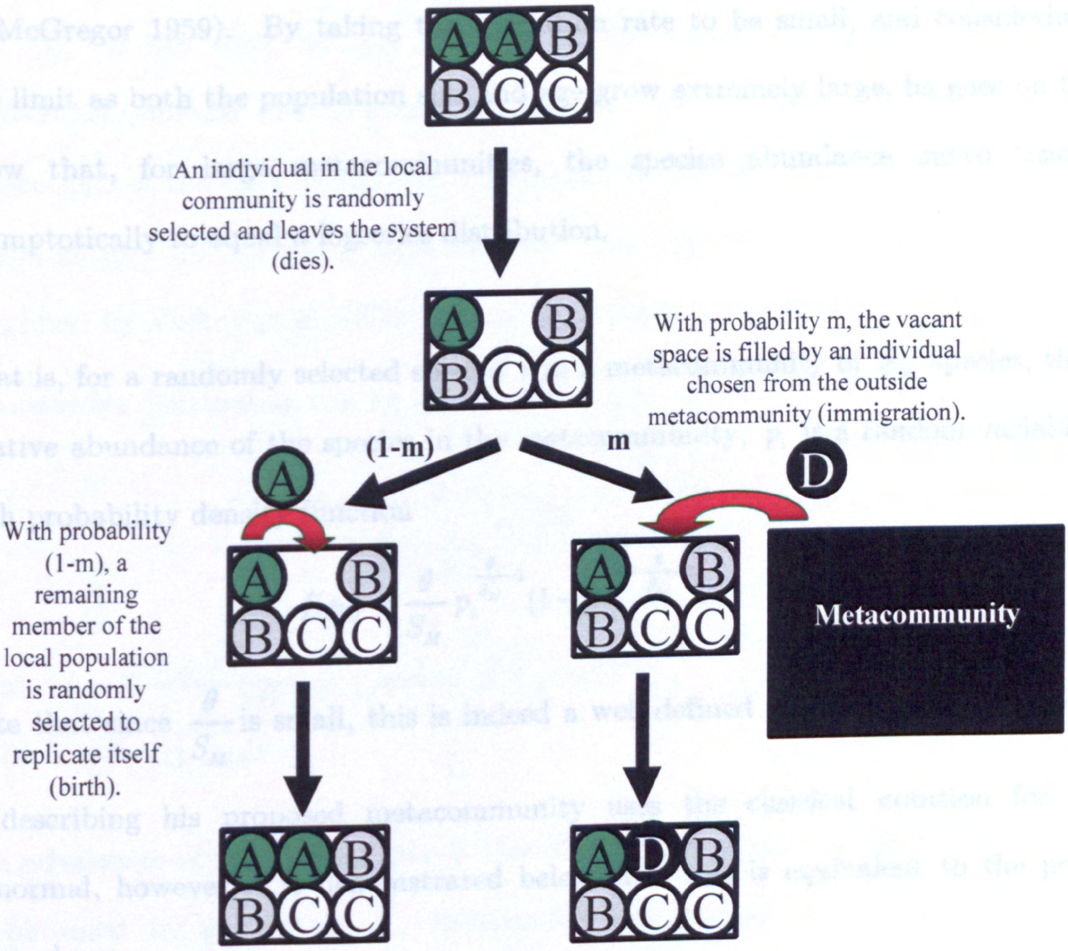
2.3.3.2 Hubbell's Neutral Community Model

One of the most recent and most hotly debated topics in ecology has been the development of Neutral Community Models (Bell 2000; Hubbell 2001). The most cited of these is the Unified Neutral theory developed by Stephen Hubbell. Through this theory he, perhaps controversially, claims to have reconciled the biogeographic inter-island predictions of MacArthur and Wilson with predictions of the relative abundances of species within each smaller local community. This is important for microbial ecology because it opens up the prospect of testing some of the concepts in the Theory of Island Biogeography, in particular the idea that local diversity is shaped as much by external ecological forces such as immigration as by local forces. Unlike MacArthur and Wilson's original theory, testing of such concepts using a neutral model becomes a feasible task, as there is no reliance upon knowledge of the total diversity, which remains immeasurable. Data can be observed only from small samples (section 2.1) which give a picture of only the right hand tail of the complete taxa-abundance distribution. As demonstrated later in this chapter and in chapter 4 of this thesis, this inability to see anything but the most common taxa is less problematic under the assumption of neutral community assembly.

The Neutral Community Model (NCM) proposed by Hubbell is built on a set of biological assumptions which are extremely basic. Certainly, any open microbial system is a birth-immigration-death process. However, in a neutral model a controversial additional assumption is made; all species at a given trophic level are assumed to be absolutely equivalent in terms of their birth and death rates. No one species is ever assumed to reproduce more rapidly, live longer or compete better for available resources than any other.

Hubbell also assumes for his model that each community is always saturated with individuals and that the total population size remains constant, fixed at this level of complete saturation. The value N_T is defined to be this fixed number of total individuals in the local population. Thus, for the structure of an assemblage to change, an individual must first die or leave the system and then its space can be occupied by either a new birth from within the system or else an immigrant from outside in order to maintain a population of size N_T.

The mechanisms governing community assembly and dynamics for a NCM are two conceptually very simple procedures (figure 2.14). Firstly, at uniform intervals in time, one individual is selected uniformly at random from the N_T within the system and is then removed from the community ('death'). To replace this organism, one of two possibilities for the second step is then chosen. With probability m, the available space and resources vacated by the dead individual are then occupied by an immigrant to the system. Immigrants are assumed to be drawn from some large outside metacommunity, which represents the set of all possible organisms which could migrate into the system. The alternative, which occurs with probability $(1-m)$, is that a remaining member of the community is selected uniformly at random and one additional member of its species is then added to the system ('birth').



An individual in the local community is randomly selected and leaves the system (dies).

With probability m, the vacant space is filled by an individual chosen from the outside metacommunity (immigration).

With probability (1-m), a remaining member of the local population is randomly selected to replicate itself (birth).

Figure 2.14
Diagram illustrating the mechanism governing community assembly for NCMs. The left hand channel shows new individuals being born into the system, the right hand illustrates immigration from the metacommunity. Immigrating individuals could be either from an entirely new species to the system (as shown here) or may also be members of a species already present within the system.

Hubbell makes one further assumption in the formulation of his NCM. Additional to the assumption of neutral dynamics, it is assumes that the diversity within the metacommunity is driven by randomly occurring speciation events. This is analogous to one of the standard models employed in population genetics (Karlin & McGregor 1959). By taking the speciation rate to be small, and considering the limit as both the population size and age grow extremely large, he goes on to show that, for large metacommunities, the species abundance curve tends asymptotically to equal a logseries distribution.

That is, for a randomly selected species i in a metacommunity of S_M species, the relative abundance of the species in the metacommunity, p_i is a random variable with probability density function

$$f(p_i) = \frac{\theta}{S_M} p_i^{\frac{\theta}{S_M}-1} (1-p_i)^{\theta-\frac{\theta}{S_M}-1}. \qquad (2.11)$$

Note that since $\frac{\theta}{S_M}$ is small, this is indeed a well-defined pdf. Hubbell's notation in describing his proposed metacommunity uses the classical notation for a lognormal, however it is demonstrated below that this is equivalent to the pdf given above.

Let $S(\mu)$ be the expected number of taxa in the metacommunity with absolute abundance μ then according to Hubbell's model of metacommunity dynamics S is described by Fisher's logseries distribution,

$$S(\mu) = \theta \frac{x^\mu}{\mu} \qquad (2.12)$$

where

$$x = 1 - e^{\left\{-\frac{S_M}{\theta}\right\}} \tag{2.13}$$

and S_M is the total number of taxa in the source community. It is not immediately obvious how to sample at random from this distribution to generate realizations of the taxa abundance distribution in the metacommunity. However, a straightforward sampling algorithm becomes apparent if an approximation suggested by Volkov et al (2003) is employed They noted that as $\theta/S_M \rightarrow 0$ then the logseries distribution can be approximated by,

$$S(\mu) = \frac{S_M}{\Gamma\left(\theta/S_M\right)\left(\frac{x}{1-x}\right)^{\theta/S_M}} e^{-\mu/\left(\frac{x}{1-x}\right)} \mu^{\theta/S_M - 1} \tag{2.14}$$

since $\dfrac{S_M}{\Gamma\left(\theta/S_M\right)\left(\frac{x}{1-x}\right)^{\theta/S_M}} \rightarrow 1$ and $e^{-\mu/\left(\frac{x}{1-x}\right)} \rightarrow e^{-\mu \ln(x)}$

The advantage of this formulation is that the species abundance distribution can be obtained by generating S_M independent realisations of Gamma variables $\mu_i \sim Gamma\left(\theta/S_M, \left(\frac{1-x}{x}\right)\right)$ for $i = 1,\ldots,S_M$ for finite θ as $\theta/S_M \rightarrow 0$.

As the variables are independent, their joint density function is simply the product of their individual density functions

$$f(\mu_1,\ldots,\mu_{S_M}) = \frac{1}{\Gamma\left(\theta/S_M\right)^{S_M}\left(\frac{x}{1-x}\right)^{\theta}} \left[e^{-\mu_1/\left(\frac{x}{1-x}\right)} \mu_1^{\theta/S_M - 1}\right] \ldots \left[e^{-\mu_{S_M}/\left(\frac{x}{1-x}\right)} \mu_{S_M}^{\theta/S_M - 1}\right] \tag{2.15}$$

However, rather than using absolute abundances which requires explicit knowledge of the number of individuals in the metacommunity, the relative abundance, p_i, of each species can be considered. Setting $p_i = {}^{\mu_i}\!\Big/\!{}_{\sum_i^{S_M} \mu_i}$ and

$N_M = \sum_1^{S_M} \mu_i$ note that only $S_M - 1$ of these p_i variables are now independent.

Therefore, set

$\mu_i = N_M p_i$ for $i = 1, \ldots, S_M - 1$

and

$\mu_{S_M} = N_M(1 - p_1 - \ldots - p_{S_M-1})$.

The joint density function of $p_1, \ldots, p_{S_M}$ is therefore

$$g(p_1, \ldots, p_{S_M-1}, N_M) = \frac{1}{\Gamma\left({}^{\theta}\!/\!{}_{S_M}\right)^{S_M} \left(\frac{x}{1-x}\right)^{\theta}} \prod_1^{S_M} \left[e^{{}^{-N_M p_i}\!/\!{}_{\left(\frac{x}{1-x}\right)}} (N_M p_i)^{{}^{\theta}\!/\!{}_{S_M} - 1} \right] \cdot |\det J| \quad (2.16)$$

where J is the Jacobian, given by,

$$\begin{pmatrix} N_M & 0 & \cdots & 0 & p_1 \\ 0 & N_M & \ddots & \vdots & p_2 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & N_M & p_{S_M-1} \\ -N_M & -N_M & \cdots & -N_M & p_{S_M} \end{pmatrix} \qquad (2.17)$$

It can be seen that $|\det J| = {N_M}^{S_M-1}$.and therefore,

$$g(\mu_1, \ldots, \mu_{S_M-1}, N_M) = \left[\frac{e^{{}^{-N_M}\!/\!{}_{\left(\frac{x}{1-x}\right)}} {N_M}^{\theta-1}}{\Gamma(\theta)\left(\frac{x}{1-x}\right)^{\theta}} \right] \left[\frac{\Gamma(\theta)\, p_1^{{}^{\theta}\!/\!{}_{S_M} - 1} \cdots p_{S_M}^{{}^{\theta}\!/\!{}_{S_M} - 1}}{\Gamma\left({}^{\theta}\!/\!{}_{S_M}\right)^{S_M}} \right] \qquad (2.18)$$

The first term of this implies that $N_M \sim Gamma\left(\theta, \left(\frac{1-x}{x}\right)\right)$, which gives that $E(N_M) = \theta\frac{x}{1-x}$ as expected for the logseries distribution. The second bracket states that $p_1, \ldots, p_{S_M}$ have a Dirichlet distribution $p_1, \ldots, p_{S_M} \sim Dir\left(\theta/_{S_M}, \ldots, \theta/_{S_M}\right)$. Additionally, $p_1, \ldots, p_{S_M}$ are independent of N_M.

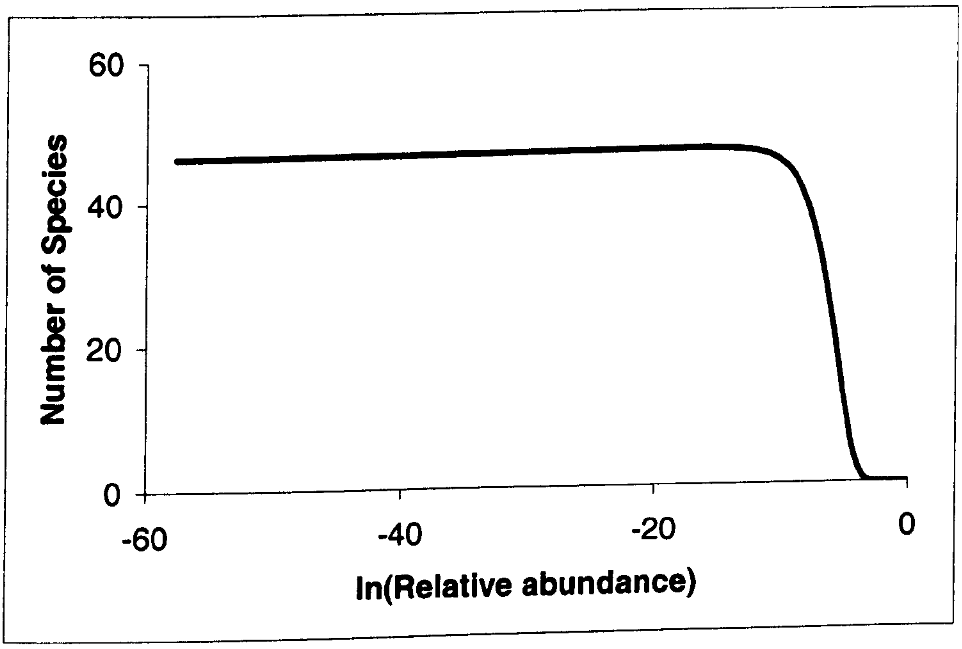


Figure 2.15

Species abundance distribution within the metacommunity for Hubbell's NCM. The diversity parameter of $\theta = 47.226$ has been found for a tree community on Barro Colorado Island, Panama (Volkov *et al.* 2003).

Mathematically, Hubbell's formulation describes a simple Markovian Urn (Grimmett & Stirzaker 2001)model. There are two different urns, one representing the individuals in the local community and one for the metacommunity. Deaths in the local system are simply selections without replacement from the local community. The species of the new member of the community is then decided by selection with replacement from one of the two urns. If the event is to be a birth, then selection is from the local community. In the case of immigration, selection is made from the metacommunity urn. An additional member of the selected species is then added to the local community. Figure 2.14 illustrates these mechanisms in a simplified system.

For any given timestep (random death followed immediately by birth or immigration from the metacommunity), the transition probabilities for N_i, the abundance of species i are

$$P\{N_i + 1 \mid N_i\} = \left(\frac{N_T - N_i}{N_T}\right)\left(mp_i + (1-m)\frac{N_i}{N_T - 1}\right) \tag{2.19}$$

$$P\{N_i - 1 \mid N_i\} = \left(\frac{N_T}{N_T}\right)\left(m(1-p_i) + (1-m)\frac{N_T - N_i + 1}{N_T - 1}\right) \tag{2.20}$$

$$P\{N_i \mid N_i\} = \\ \left(\frac{N_i}{N_T}\right)\left(mp_i + (1-m)\frac{N_i - 1}{N_T - 1}\right) + \left(\frac{N_T - N_i}{N_T}\right)\left(m(1-p_i) + (1-m)\frac{N_T - 1}{N_T - 1}\right) \tag{2.21}$$

Each of these is easily apparent from the urn model definition. Take, for example equation(2.19). This gives the probability of the abundance of the species in question, species i, increasing during any given timestep. For this to happen, a member of any species other than species i must die. This occurs with probability$\left(\frac{N_T - N_i}{N_T}\right)$, since that is the proportion of individuals in the systems which are not of the species in question. Then, in order for the species abundance to increase, one of two events may occur. Either an immigrant of species iI s chosen from the metacommunity, which occurs with probability mp_i, or else one of the remaining individuals is chosen to give birth to an identical individual, which occurs with probability$(1-m)\frac{N_i}{N_T - 1}$. Multiplying the death probability by the sum of the birth and immigration probabilities gives equation (2.19). Equations (2.20) and(2.21) can be similarly justified by considering what must happen during the timestep for a given species abundance to decrease or remain stationary.

Finding the steady state distribution from these transition probabilities is not a trivial task. Until recently, the distribution tended to have been obtained from simulations, although there have been recently published exact analytic forms (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003; McKane *et al.* 2004). Both of these approaches reveal that for high immigration systems, the local community tends to resemble the metacommunity, and appear logseries-like. As local migration drops, the total diversity in the system decreases and the left hand tail of the species abundance curve declines, making the curve appear more lognormal-like. As this migration rate continues to decline, the biodiversity drops ever further and the curve becomes increasingly left skewed with the rarer species the first to be lost. In the extreme limit of migration tending towards zero, immigration collapses completely, and one single species becomes monodominant.

Figures 2.16 and 2.17 below respectively illustrate the species abundance predicted by Hubbell's NCM for high immigration and low immigration systems.

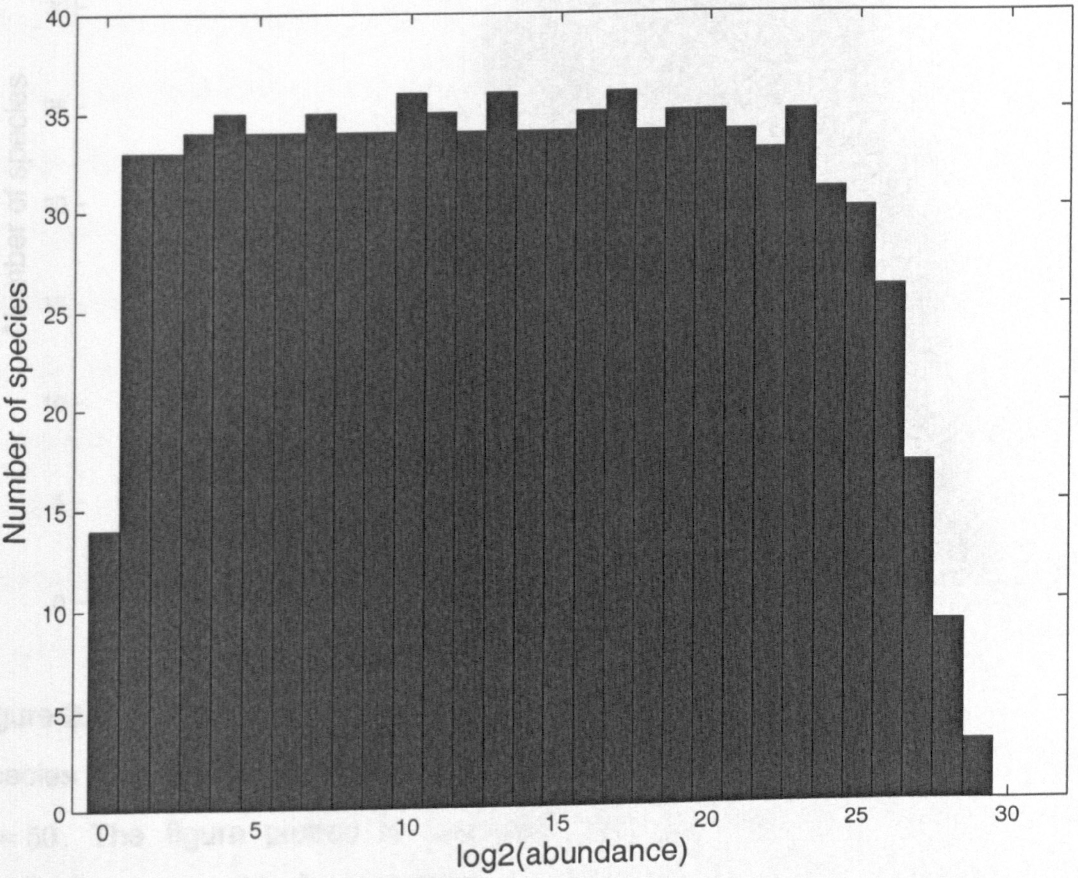


Figure 2.16
Species abundance distribution for a NCM where $N_T = 10^{10}$, $m = 1$ and $\theta = 50$. The figure plotted is averaged over 100 repetitions. Note that the distribution closely resembles that of the metacommunity, which has a logseries distribution.

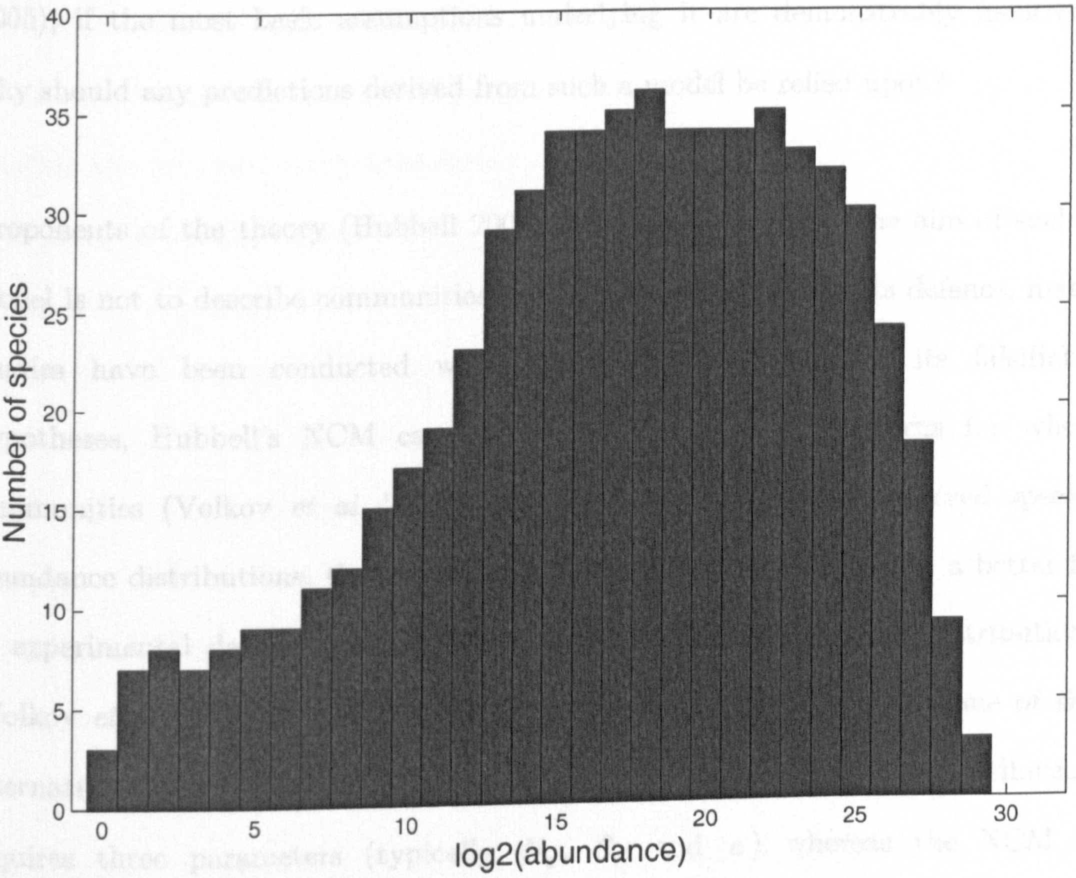


Figure 2.17

Species abundance distribution for a NCM where $N_T = 10^{10}$, $m = 10^{-4}$ and $\theta = 50$. The figure plotted is averaged over 100 repetitions. Note that the distribution more closely resembles a lognormal distribution than that of the metacommunity.

2.3.3.3 The Neutrality Contoversy

Neutral models are always, by their very nature, controversial. Undeniably, examination of microbial systems at the individual level will reveal that the birth and death rates of all taxa are most certainly not equivalent. This has led many critics to question the worth of Hubbell's neutral model (McGill 2003; Wootton

2005); if the most basic assumptions underlying it are demonstrably incorrect, why should any predictions derived from such a model be relied upon?

Proponents of the theory (Hubbell 2001) argue, however, that the aim of such a model is not to describe communities at the individual scale. In its defence, many studies have been conducted which show that, in spite of its falsifiable hypotheses, Hubbell's NCM can indeed explain diversity patterns for whole communities (Volkov *et al.* 2003). By calibrating a NCM to observed species abundance distributions, the neutral model has often been seen to be a better fit to experimental datasets than some of the other commonly cited distributions (Volkov *et al.* 2003), despite having fewer degrees of freedom than some of the alternative theories. For example, the lognormal species abundance distribution requires three parameters (typically N_0, S_T and a) whereas the NCM is dependent only upon θ and the combined $N_T m$ parameter.

The inaccuracies at the individual scale, it is argued, should not lead to NCMs being dismissed completely, given their abilities to offer explanations and predictions at the entire community scale (Bell 2000; Enquist *et al.* 2002). Many fields of the natural sciences employ theories derived from assumptions which are, on some scale, palpably incorrect. Certain factors can often be ignored which, although sizeable at the point scale, soon cancel out and appear negligible if the entire system is the topic of interest.

2005); if the most basic assumptions underlying it are demonstrably incorrect, why should any predictions derived from such a model be relied upon?

Proponents of the theory (Hubbell 2001) argue, however, that the aim of such a model is not to describe communities at the individual scale. In its defence, many studies have been conducted which show that, in spite of its falsifiable hypotheses, Hubbell's NCM can indeed explain diversity patterns for whole communities (Volkov *et al.* 2003). By calibrating a NCM to observed species abundance distributions, the neutral model has often been seen to be a better fit to experimental datasets than some of the other commonly cited distributions (Volkov *et al.* 2003), despite having fewer degrees of freedom than some of the alternative theories. For example, the lognormal species abundance distribution requires three parameters (typically N_0, S_T and a) whereas the NCM is dependent only upon θ and the combined $N_T m$ parameter.

The inaccuracies at the individual scale, it is argued, should not lead to NCMs being dismissed completely, given their abilities to offer explanations and predictions at the entire community scale (Bell 2000; Enquist *et al.* 2002). Many fields of the natural sciences employ theories derived from assumptions which are, on some scale, palpably incorrect. Certain factors can often be ignored which, although sizeable at the point scale, soon cancel out and appear negligible if the entire system is the topic of interest.

One of the most celebrated of these theories to arise from falsifiable hypotheses is the Ideal Gas Law in physics. It remains the most commonly utilised equation relating the pressure, temperature and volume of gases. However, its formulation is built upon assumptions which, at least at the scale of the individual molecules of the gas, are easily shown to be extremely inaccurate. The assumptions are made that all collisions between molecules are completely elastic, and any other interactions between the molecules can be considered to be negligible. If employed to studies of miniscule volume of gas, these errors in the formulation can prove sizeable. However, for systems of greater size all these small errors at the point scale are seen to cancel out and the result of this set of overly simplistic assumptions is the celebrated equation

$$pV = nRT$$

which proves to be extremely accurate for almost all practical studies. Despite the incorrect assumptions relied upon in its derivation, this remains one of the most commonly employed equations in the natural sciences, and one which has proven extremely accurate when applied at anything larger than the point scale.

That said, neutrality is a concept with a great number of critics (McGill 2003). For example, many commonly cited models assume that the major factor in shaping microbial systems is niche construction (Laland *et al.* 1999; Torsvik *et al.* 2002; Condit *et al.* 2006). These alternative models are based on the premise that the most predominant mechanism in shaping community assembly is the ability of species to alter their habitat to their own benefit. Many studies have demonstrated that, within small localised regions, organisms are often able to affect the physical and chemical conditions within their habitat to make them advantageous to their own survival and to maximise the growth rate of their own species.

Clearly, such a theory opposes the most fundamental premise of NCMs. The central assumption of any neutral model holds that on the scale of large communities, for which NCMs can be applied, the net effects of any such small local competitive advantages or disadvantages are negligible when the whole system is considered. While the assumption of neutrality does not directly contradict the niche assembly findings at the individuals level, it does maintain that any such increases or decreases in species fitness within niches soon cancel out. That is, even if, a species has an advantage in terms of a higher growth rate than others within one niche, there will exist other niches at which it is at a competitive disadvantage.

2.3.4 Other Ecological Models

Although this literature review has considered many of the ecological models which are relevant to the research presented later in this thesis, it by no means provides an exhaustive overview of ecological theory. Indeed there are huge areas of the field which have not been reviewed, for example resource-ratio theory (Tilman 1976), predator-prey models (Volterra 1931)and food webs (Sugihara *et al.* 2003). In fact, resource competition models have already played a central role in waste treatment technologies. The pair of differential equations (section 2.1.2) used in design essentially describe a resource competition model but with functional groups as the taxonomic resolution rather than the individual species.

Ultimately, environmental engineers should aspire to understand how all of these ecological processes affect the community composition and functioning of microbial communities. However, most theories in their current form require a degree of specificity in growth kinetics, niche affiliations and life histories that is beyond the reach of microbial ecologists employing current environmental technologies.

There are two significant attractions in investigating how neutral theories might be applied to microbial systems. Firstly, they describe the community assembly process. Several reviews and perspectives on theory in biotechnologies (Rittmann *et al.* 2006) hold up a quantitative description of microbial community assembly as a major goal. Secondly, neutral theories are parsimonious description of reality and as such have few parameters. Thus it is conceivable that, if they hold, they can be parameterised and tested. Both the parameterisation and testing of neutral theories are considered later on in this thesis.

3 Neutral Community Models for Microbial Systems

In the previous chapter of this thesis, Neutral Community Models (NCMs) were introduced, particularly the NCM developed by Stephen Hubbell in his attempt to unify two theories on biogeography and biodiversity. For studies in classical ecology NCMs have been shown to replicate many of the fundamental patterns observed in many biological communities, from insects to trees (Hubbell 2001), such as taxa-abundance distributions and taxa-area relationships observed in many systems. These relationships are seen as some of the central tenets of community ecology (Levin 2000; Green *et al.* 2004), and ecologists have been seeking theoretical explanations for their formation for over a century.

The success of NCMs is remarkable, given that they assume nothing beyond a simple birth-death-immigration process but, as discussed in section 2.3, these models have been met with scepticism by many ecologists (McGill 2003; Wootton 2005). Whilst the proponents of NCMs point to their success in explaining observed patterns, the sceptics suggest that the underlying ecological mechanisms remain unproven (McGill 2003). Later in this thesis, some of these criticisms are addressed.

In this thesis, a NCM will be used as the basis for modelling wastewater treatment systems. As discussed in the previous chapter, the goal of this thesis is not to directly approach any single engineering problem, rather to lay the groundwork for a fuller understanding of the mechanisms by which microbial systems assemble and then alter through time. Indisputably, wastewater treatment systems are open systems, constantly subject to invasion events from other microbes, whether air-borne or within the influent. By conceiving the metacommunity as the distribution of individuals within these two sources which could survive and reproduce within the treatment plant, the basic mechanisms of the NCM can be applied to these systems.

Additionally, such a model allows for the examination of the effects of immigration in isolation of other factors. There is significant evidence (Hubbell 2001) that even rare immigration events can be the driving force behind shaping community structure in large communities. Given the current design of such plants (section 2.1) immigration, either by increased mixing or quicker throughput, is something which could be relatively simply manipulated.

At the moment, the spatial scales at which the neutral model may apply remain unknown. Should the whole treatment plant be considered the local community? Should the plant be conceived as many interacting local communities? At the moment, any internal spatial structure of these communities has been neglected and the whole system pictured as the local community. This may well need to be amended and improved in the near future (chapter 9). It is also accepted that, at some time in the future, more complexities may be required to fully model such systems (such as inter-species competition), a neutral model will at very least serve as a good null model against which to test other hypotheses.

Previously published Neutral Community models are discrete; each birth, death or immigration is explicitly represented. This makes them prohibitively computationally inefficient for microbial communities (McGill 2003) where in for example a millilitre of wastewater there can be as many as 10^9 organisms (Whitman *et al.* 1998) and in full scale activated sludge plant up to 10^{18}. Not only this but in such large populations it is unlikely that one would ever be interested in whether the population of a particular species increases or decreases by one individual. From the environmental engineer's viewpoint, system failure occurs when the microbial communities fluctuate by an order or magnitude or more. For example, if the system is inefficient with a million ammonia oxidising bacteria, knowing that a birth has just increased that population to a million and one will scarcely improve plant function. Accordingly, some accuracy can be sacrificed so that communities of the sizes equivalent to those of microbial systems can be modelled.

Furthermore, it may only be the long-term steady-state distribution of taxa-abundances or species richness that is of interest. In which case simulating each discrete demographic event as the community evolves towards a steady-state, as Hubbell (2001) and Bell (2000) do, is highly inefficient and an alternative formulation of the model from which a steady states can be found directly is highly desirable. Despite such models being reliant upon a very simple set of assumptions and quite straightforward mathematical formulation, they are nonetheless inapplicable to microbial communities. Furthermore, explicitly representing each birth, death or immigration in any system proves to be a redundant exercise if the main interest is in overall diversity. Thus the discrete form of Hubbell's NCM cannot be applied to very large biological communities and the model must be reformulated for application to microorganisms.

In this chapter, Hubbell's NCM is modified and further developed to produce a new model that can predict the assembly and composition of arbitrarily large communities. This allows for the application of a NCM to large microbial systems such as those encountered in typical wastewater treatment systems. The new NCM developed in this chapter is utilised extensively in the remainder of the thesis to explore a variety of features of microbial community assembly.

The main research achievements in this chapter include:

- Derivation of a multidimensional diffusion equation for the probability density function of all species abundances within a microbial population, via approximation of the transition probabilities in Hubbell's Markovian matrix.
- Solution of the diffusion equation for when the local community is in steady-state. This provides a simple tool for simulating such communities which is vastly quicker and computationally simpler than any other method currently employed.
- Formulation of a simple approach for estimating immigration rates into similar microbial systems from simple presence-absence datasets.

These achievements form the basis of a publication in Environmental Microbiology (Sloan et. al 2006)

3.1 Derivation of a Diffusion Equation for the Probability Density Function of All Species

The derivation of a new formulation for the Neutral Community Model stems from the simple assumption that, given the microbial populations of interest to engineers tend to be very large, the abundances of each taxon can be assumed to be continuous variable.

However, to start with, consider Hubbell's discrete model written in a slightly different way. Rather than formulating it as a discrete Markov chain it can be written as a simple birth-immigration death process with a fixed population size of N_T. The probabilities of the abundance of taxon i growing, decreasing or staying stationary during the Rth timestep are just the transition probabilities in Hubbell's model. Accordingly, when its abundance $N_i = k$ the probability of taxon i's population growing by one individual is

$$G_k(R) = \left(\frac{N_T - k}{N_T}\right)\left(mp_i + (1-m)\frac{k}{N_T - 1}\right) \tag{3.1}$$

the probability of it decreasing by one is

$$D_k(R) = \left(\frac{k}{N_T}\right)\left(m(1-p_i) + (1-m)\frac{N_T - k + 1}{N_T - 1}\right) \tag{3.2}$$

and the probability of it staying the same is

$$\begin{aligned} S_k(R) = &\left(\frac{k}{N_T}\right)\left(mp_i + (1-m)\frac{k-1}{N_T - 1}\right) \\ &+ \left(\frac{N_T - k}{N_T}\right)\left(m(1-p_i) + (1-m)\frac{N_T - k + 1}{N_T - 1}\right) \end{aligned} \tag{3.3}$$

Now, for a taxon to have k individuals present at a given time, one of three events must have occurred since the previous timestep: the taxon had $k-1$ individuals and one additional one was added, it had $k+1$ individuals and one was removed, or it had k individuals and there was no net change in its abundance. So let $P_k(R)$ be the probability that a particular taxon of interest has abundance $N_i = k$ in the local community after R timesteps. $P_k(R)$ can then be expressed conditionally as a function of all its possible states one timestep previously, giving the following one-dimensional difference equation

$$P_k(R+1) = G_{k-1}(R)P_{k-1}(R) + D_{k+1}(R)P_{k+1}(R) + (1 - G_k(R) - D_k(R))P_k(R)$$

which gives

$$\begin{aligned} &P_k(R+1) = \\ &P_k(R) + \left[D_{k+1}(R)P_{k+1}(R) - D_k(R)P_k(R)\right] + \left[G_{k-1}(R)P_{k-1}(R) - G_k(R)P_k(R)\right] \end{aligned}. \quad (3.4)$$

However, in order to extend the scope of the NCMs to cover the very large population sizes observed for microbial systems, then an analogous result to equation(3.4) is required, one which allows for the number of individuals in the system N_T to be made arbitrarily large.

First, a change of variable is required. Instead of looking simply at the absolute abundance of the ith taxon, its relative abundance is examined:

$$x_i = \frac{N_i}{N_T}$$

where N_T is the total number of individual organisms in the community and hence $N_T = \sum N_i$. In Hubbell's model and this extension of it, the environment is assumed to be saturated with individuals and, therefore, N_T is constant through time. Here, unlike in Hubbell's formulation, it can be arbitrarily large. In the limit for large values of N_T, x_i is taken to be a continuous variable with domain $[0,1]$.

Additionally, rather than considering the number of timesteps taken, N, a continuous time variable, t, is required. As formulated in the discrete case and the difference equation, changes in the population occur at timesteps which occur at uniform time intervals. However, assuming a fixed specific death rate as Hubbell does, the length between each timestep must be proportional to $1/N_T$ since, for example, doubling the population size will double the total death rate in the population and thus halve the time between each transition event in the system. Thus, for extremely large values of N_T, the time between successive death-replacement events will be greatly reduced.

This change to continuous relative abundance and time variables means that instead of looking for individual probabilities that the population is in each possible state, a probability density function (pdf) $\phi(x_i,t)$ is sought for the chance that the abundance of the species at time t is x_i.

Now, define $D(x_i)$ and $G(x_i)$ in the same way as for the discrete case to be the respective probabilities of the abundance of taxon i decreasing or growing during any unit interval of time. Note that both these quantities are dependent only upon the abundance variable x_i and are independent of t.

The continuous equivalent for the difference equation (3.4) is then

$$\begin{aligned}\phi(x_i,t+\delta t) = \phi(x_i,t)&\\ &+\left[D(x_i+\delta x)\phi(x_i+\delta x_i,t)-D(x_i)\phi(x_i,t)\right]\\ &+\left[G(x_i-\delta x)\phi(x_i-\delta x_i,t)-G(x_i)\phi(x_i,t)\right]\end{aligned} \tag{3.5}$$

Taylor expansions can then be applied to get

$$\phi(x_i,t+\delta t)\approx\phi(x_i,t)+\delta t\frac{\partial\phi}{\partial t}(x_i,t)+\ldots \tag{3.6}$$

$$\begin{aligned}D(x_i+\delta x_i)\phi(x_i+\delta x_i,t)\approx D(x_i)\phi(x_i,t)&\\ &+(\delta x_i)\frac{\partial\left[D(x_i)\phi(x_i,t)\right]}{\partial x_i}\\ &+\frac{1}{2}(\delta x_i)^2\frac{\partial^2\left[D(x_i)\phi(x_i,t)\right]}{\partial x_i^{\,2}}+\ldots\end{aligned} \tag{3.7}$$

and

$$\begin{aligned}G(x_i-\delta x_i)\phi(x_i-\delta x_i,t)\approx G(x_i)\phi(x_i,t)&\\ &-(\delta x_i)\frac{\partial\left[G(x_i)\phi(x_i,t)\right]}{\partial x_i}\\ &+\frac{1}{2}(\delta x_i)^2\frac{\partial^2\left[G(x_i)\phi(x_i,t)\right]}{\partial x_i^{\,2}}-\ldots\end{aligned} \tag{3.8}$$

Ignoring higher order terms, substituting these three expansions into equation (3.5) and letting the timestep $\delta t\to 0$ gives the following one-dimensional Fokker-Planck equation for the pdf $\phi(x_i,t)$of the abundance x_i at time$t$:

$$\frac{\partial\phi(x_i,t)}{\partial t}=-\frac{\partial(M_{\delta x_i}\phi(x_i,t))}{\partial x_i}+\frac{1}{2}\frac{\partial^2(V_{\delta x_i}\phi(x_i,t))}{\partial x_i^2} \tag{3.9}$$

where the quantities $M_{\delta x_i}$ and$V_{\delta x_i}$ are, defined as

$$M_{\delta x_i}=\lim_{\delta t\to 0}\frac{\delta x_i\left[G(x_i)-D(x_i)\right]}{\delta t}$$

$$V_{\delta x_i}=\lim_{\delta t\to 0}\frac{\delta x_i^{\,2}\left[G(x_i)+D(x_i)\right]}{\delta t}$$

Now, with the change to continuous variables, the changes in the relative abundances per unit time are always of magnitude $1/N_T$. To obtain estimates of $M_{\delta x_i}$ and $V_{\delta x_i}$, it is assumed that the difference equation (3.4) is a suitable approximation to the diffusion equation. In the discrete case $\delta x_i/\delta t$ and $\delta x_i^2/\delta t$ are, respectively, $1/N_T$ and $1/N_T^2$ since the timesteps are of unit length, thus $\delta t = 1$.

Accordingly, the quantities $M_{\delta x_i}$ and $V_{\delta x_i}$ are, respectively takes as,

$$M_{\delta x_i} = \frac{G_k - D_k}{N_T}$$

$$V_{\delta x_i} = \frac{G_k + D_k}{N_T^2}$$

Working with the continuous variables, and N_T be sufficiently large such that $N_T - 1 \approx N_T$, the three transition functions for when $x_i = k / N_T$ ((3.1)(3.2) and (3.3)) tend to the following quantities

$$G_k = (1 - x_i)[mp_i + (1 - m)x_i] \tag{3.10}$$

$$D_k = x_i[m(1 - p_i) + (1 - m)(1 - x_i)] \tag{3.11}$$

$$S_k = x_i[mp_i + (1 - m)x_i] + (1 - x_i)[m(1 - p_i) + (1 - m)(1 - x_i)]. \tag{3.12}$$

which in turn give

$$M_{\delta x_i} = \frac{m(p_i - x_i)}{N_T} \tag{3.13}$$

$$V_{\delta x_i} = \frac{2x_i(1 - x_i) + m(p_i - x_i)(1 - 2x_i)}{N_T^2}. \tag{3.14}$$

However, in practice, it is seen that m is typically very small, allowing for the second term in the expression for $V_{\delta x_i}$ to be considered negligible.

Even in the theoretical case where m is unusually large, x_i would rapidly converge towards p_i thus vanishing the second term. Consequently, equation (3.14) is taken as

$$V_{\delta x_i} \approx \frac{2x_i(1-x_i)}{N_T^2} \tag{3.15}$$

The procedure described above can be generalised to give a higher dimensional equivalent to this diffusion equation that describes the joint probability density function for the abundance of all taxa in the community, rather than just one taxon. Once more, the starting point is a difference equation.

Let $P_K(n)$ be the probability that the abundances of all n taxa are $K=(k_1,k_2,\ldots,k_n)$.

Then, define $P_K^{i^+}(n)$ to be the probability that abundance vector is $(k_1,\ldots k_i+1,\ldots,k_n)$ and $P_K^{i^-}(n)$ to be the probability that the abundances are $(k_1,\ldots k_i-1,\ldots,k_n)$, and set $G_K^{i^-}$ $G_K^{i^+}$ $D_K^{i^-}$ $D_K^{i^+}$ similarly for the transition probabilities. Considering the current state of the system as a function of all possible states it may have been in one timestep previous (as with equation(3.4)) results in the following difference equation

$$\begin{aligned} P_K(N+1) &= \sum_{i=1}^{n}\left[G_K^{i^-}(N)P_K^{i^-}(N)\right]+\sum_{i=1}^{n}\left[D_K^{i^+}(N)P_K^{i^+}(N)\right] \\ &+\sum_{i=1}^{n}\left[(1-G_K^{i}(N)-D_K^{i}(N))P_K^{i}(N)\right]+\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\left[X_K^{i^-,j^+}P_K^{i^-,j^+}(N)\right] \end{aligned} \tag{3.16}$$

where $X_K^{i^-,j^+}$ is the probability that the abundance vector changes from$=(k_1,...k_i-1,...,k_j+1,...,k_n)to(k_1,...,k_n)$ in a given timestep.

Again, Taylor expansions are required to simplify this expression in the continuum limit, and the result is again a Fokker-Planck equation, only this time in $(n-1)$ dimensions.

$$\frac{\partial\phi}{\partial t}=\sum_{i=1}^{n}\left[-\frac{\partial(M_{\delta x_i}\phi)}{\partial x_i}+\frac{1}{2}\frac{\partial^2(V_{\delta x_i}\phi)}{\partial x_i^2}\right]+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}\frac{1}{2}\frac{\partial^2(C_{\delta x_i\delta x_j}\phi)}{\partial x_i\partial x_j} \tag{3.17}$$

where $M_{\delta x_i}$, $V_{\delta x_i}$ and $C_{\delta x_i\delta x_j}$ are

$$M_{\delta x_i}=\lim_{\delta t\to 0}\frac{\delta x_i\left[G(x_i)-D(x_i)\right]}{\delta t}$$

$$V_{\delta x_i}=\lim_{\delta t\to 0}\frac{\delta x_i^{\,2}\left[G(x_i)+D(x_i)\right]}{\delta t}$$

$$C_{\delta x_i\delta x_i}=\lim_{\delta t\to 0}\frac{\delta x_i\delta x_j\left[X^{i^-,j^+}+X^{j^-,i^+}\right]}{\delta t}.$$

For the NCMs being considered, these equate to

$$M_{\delta x_i}=\frac{m\left(p_i-x_i\right)}{N_T}$$

$$V_{\delta x_i}=\frac{2x_{\;i}\left(1-x_i\right)+m\left(p_i-x_i\right)\left(1-2x_i\right)}{N_T^2}\approx\frac{2x_{\;i}\left(1-x_i\right)}{N_T^2}$$

$$C_{\delta x_i x_j}=-\left[\frac{2x_{\,i}x_j+m[x_i(p_j-x_j)+x_j(x_i-p_i)]}{{N_T}^2}\right]\approx-\left[\frac{2x_ix_j}{{N_T}^2}\right]$$

where again, smaller order terms are considered negligible because of the typically small magnitude of m or rapid convergence of x_i to p_i.

The diffusion equations (3.9) and (3.17), allow the investigation of both the steady-state probability density function, upon which the distribution of taxa abundances converges after a suitably long timespan (Section 3.2), and the transient dynamics of populations through time. This latter topic of study is briefly discussed in this thesis as an interesting line of possible future research.

3.2 Equilibrium Species Abundances Predicted by the Adapted NCM

The original application of Hubbell's neutral community model was to explain the diversity patterns observed in communities of larger organisms, such as tree and bird populations. The vast majority of extensive ecological surveys consist of census data at one or more sites of interest at one given time. Particularly for slowly changing systems, such as those of tree communities or other organisms with long lifespans, tracking the communities composition through time can take many years or even decades to gather. Whilst such datasets do exist, for example the extensive survey of red deer on the Isle of Rum (Clutton-Brock *et al.* 1997), financial and time constraints make time series data of whole communities extremely rare. The problems faced in producing such datasets for microbial systems are somewhat different. At least in theory, it is feasible to collect ecologically relevant time series data over realistic time periods because the growth rates in the populations are so much higher. However, time series of community composition in natural microbial communities are scarce because generating quantitative data using current molecular methods has been laborious. This may change when high throughput molecular methods, such as those being pioneered by Sogin *et al.* (2006) are optimised and become routinely available.

Thus, in all previous studies where ecologists have attempted to test NCMs they have done so by assuming that the community composition is in a steady state and thus they compare observed distribution of taxa with predicted steady-state, equilibrium distributions.

Accordingly, steady-state solutions of the NCMs that give the equilibrium taxa-abundance distribution have been sought. There are a few highly cited papers on NCMs that have focused on deriving analytical solutions to Hubbell's NCM, such as Vallade & Houchmandzadeh (2003) and Volkov *et al.* (2003). These are, however, extremely difficult and slow to apply to real systems, particularly those containing vast population sizes. For example, the solution derived by Vallade & Houchmandzadeh (2003) is dependent upon the calculation of factorials of the order of N_T and Volkov *et al.* (2003) requires a complex numerical optimisation routine for a somewhat convoluted integral. For microbial systems of the order of 10^{18} individuals, it soon becomes apparent that neither of these approaches offers solutions which prove to be computable.

The approach here is to derive the steady-state distribution directly from the Fokker-Planck equations (equations (3.9) and(3.17)). These were themselves formulated to simulate the action of a large number of molecules on a similar scale to that of microbial systems, and thus the resulting distribution can predict the composition of communities of any size. Rather than relying upon the calculation of massive factorials, an equilibrium distribution can be found which is reliant solely upon simple and computationally quick mathematical functions which can be generated with any basic mathematical software package.

3.2.1 One-Dimensional Steady-State Distribution

Initially, a solution to the simple one-dimensional Fokker-Planck equation is developed. This provides the marginal distribution for the abundance attained by just one taxon embedded within a community that is undergoing neutral dynamics, irrespective of the abundances of the others. Knowledge of this distribution is particularly useful in elucidating some of the properties of a neutral community. In addition, the marginal distributions for taxa are used in developing a novel way of calibrating neutral models (Section 3.3) which is necessary for microbial communities where partial taxa-abundance data has been obtained using molecular methods.

Before any solutions for the Fokker-Planck equations, whether the single- or multi-dimensional case, can be sought, the boundary conditions must be established.

First consider the simpler one-dimensional case. The solution for ϕ must be a pdf with domain $[0,1]$. Therefore,

$$\int_{-\infty}^{0} \phi dx_i = 0 \tag{3.18}$$

$$\int_{0}^{1} \phi dx_i = 1 \tag{3.19}$$

Taking equation (3.18)first and differentiating

$$\frac{\partial}{\partial t}\int_{-\infty}^{0} \phi dx_i = \int_{-\infty}^{0} \frac{\partial \phi}{\partial t} dx_i = 0 \quad .$$

Substituting equation (3.9) then gives

$$0 = \int_{-\infty}^{0} \frac{\partial \phi}{\partial t} dx_i = \int_{-\infty}^{0} \frac{\partial}{\partial x_i}\left[-M_{\delta x_i}\phi + \frac{1}{2}\frac{\partial (V_{\delta x_i}\phi)}{\partial x_i}\right] dx_i$$

thus

$$\left[-M_{\delta x_i}\phi + \frac{1}{2}\frac{\partial (V_{\delta x_i}\phi)}{\partial x_i}\right] = 0 \text{ at } x_i = 0 \tag{3.20}$$

Similarly, equation (3.19)implies that

$$\left[-M_{\delta x_i}\phi + \frac{1}{2}\frac{\partial (V_{\delta x_i}\phi)}{\partial x_i}\right] = 0 \text{ at } x_i = 1 \tag{3.21}$$

Equations (3.20) and (3.21) define the boundary conditions. They imply that $x_i = 0$ and $x_i = 1$ can be considered to be constant flux boundaries. With these conditions, an analytic solution can be obtained directly from the one-dimensional Fokker-Planck equation.

Firstly, the time differential in the Fokker-Planck equation (3.9) is set to zero and both sides are integrated with respect to x_i.

$$const. = -M_{\delta x_i}\phi + \frac{1}{2}\frac{\partial (V_{\delta x_i}\phi)}{\partial x_i} \,. \tag{3.22}$$

However, the boundary conditions imply this constant is necessarily zero, since it is zero at both $x_i = 0$ and $x_i = 1$. Hence

$$\frac{1}{V_{\delta x_i}\phi}\frac{\partial(V_{\delta x_i}\phi)}{\partial x_i} = 2\frac{M_{\delta x_i}}{V_{\delta x_i}},$$

and therefore,

$$\begin{aligned}\phi &\propto \frac{1}{V_{\delta x}}\exp\left[2\int\frac{M_{\delta x}}{V_{\delta x}}dx_i\right]\\ &= \frac{1}{x_i(1-x_i)}\left[\int\left[\frac{N_T m p_i}{x_i(1-x_i)} - \frac{N_T m x_i}{x_i(1-x_i)}\right]dx_i\right].\\ &= \frac{x_i^{N_T m p_i}\left(1-x_i\right)^{N_T m(1-p_i)}}{x_i(1-x_i)}\end{aligned}$$

subject to the constraint that $\int_0^1 \phi dx_i = 1$ (from equations (3.18) and (3.19)), then

$$\phi = \frac{\Gamma(N_T m)}{\Gamma(N_T m p_i)\Gamma(N_T m(1-p_i))}x_i^{N_T m p_i - 1}\left(1-x_i\right)^{N_T m(1-p_i)-1} \tag{3.23}$$

which means that the relative population size $x_i \sim Beta[N_T m p_i, N_T m(1-p_i)]$.

From this, the expected abundances of each taxon and their variances can readily be obtained.

$$\begin{aligned}E[x_i] &= \left[\frac{\Gamma(N_T m)}{\Gamma(N_T m p_i)\Gamma(N_T m(1-p_i))}\right]\cdot\left[\frac{\Gamma(N_T m p_i + 1)\Gamma(N_T m(1-p_i))}{\Gamma(N_T m + 1)}\right]\\ &= \left[\frac{N_T m p_i}{N_T m}\right] = p_i\end{aligned} \tag{3.24}$$

$$
\begin{aligned}
Var[x_i] &= \left[\frac{\Gamma(N_T m)}{\Gamma(N_T m p_i)\Gamma(N_T m(1-p_i))}\right]\cdot\left[\frac{\Gamma(N_T m p_i+2)\Gamma(N_T m(1-p_i))}{\Gamma(N_T m+2)}\right]-p_i^{\,2} \\
&= \left[\frac{(N_T m p_i+1)N_T m p_i}{(N_T m+1)N_T m}\right]-p_i^{\,2} \qquad (3.25) \\
&= \left[\frac{p_i(1-p_i)}{N_T m+1}\right]
\end{aligned}
$$

The Beta distribution represents a family of curves which, depending upon its parameters could take a variety of different shapes. Mathematically, Beta distributions can have a mode of 1, be bi-modal with modes at both 0 and 1 or even have a uniform distribution on the domain$[0,1]$. However, the biological conditions required to produce such pdfs are extremely unlikely to arise. For example, the uniform distribution can only occur when $p_i = \frac{1}{2}$ and $N_T m = 1$.

For the parameter ranges one might anticipate for real microbial systems, one of two very different types of curve is likely to represent the pdf of a taxon's abundance within a community, with the $N_T m$ parameter as the determining factor as to which of these is the case. For lower immigration systems, the distribution appears hyperbolic-like, with a modal value of 0. As immigration increases, an internal mode arises with the distribution forming a skewed bell curve centred around the modal value of$\frac{N_T m p_i - 1}{N_T m - 2}$. Given that the variance of $x_i$is inversely proportional to$N_T m$ (equation(3.25)), it is easy to see that for large values of$N_T m$, the pdf forms a tight bell curve around the value ofp_i.

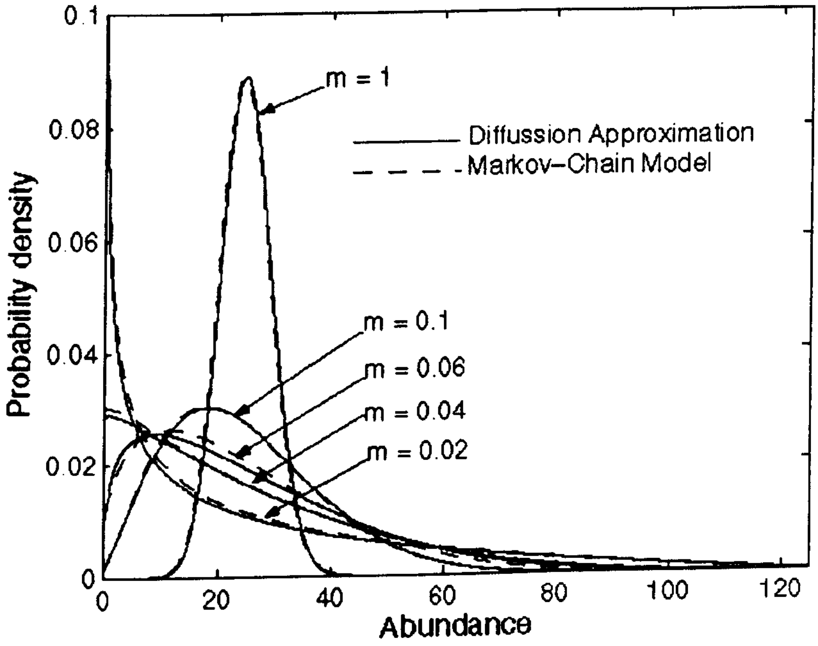


Figure 3.1

A comparison of stationary distributions for a variety of immigration rates, using both Hubbell's Markov model and the continuous variant developed here. Here $p_i = 0.2$.

Figure 3.1 illustrates two notable phenomena. Firstly, even for very small populations (in this case 125 individuals) the approximation via the diffusion equation yields very similar results to Hubbell's original Markov Chain model. The second significant feature is that for a fixed population size, it is seen that the variance of the pdf increases as the immigration rate drops. For high immigration systems, there is an interior mode for the pdf. For a fixed community size, there is a threshold for immigration below which the mode of the distribution goes to zero.

It is desirable to know the exact point of the threshold at which the mode of the pdf ϕ becomes zero. Beyond this threshold, the species is most likely to be completely absent from any community. This criterion for judging whether or not a stationary distribution is desirable from the environmental engineer's viewpoint may seem a bit arbitrary. In practice, measurement techniques can only detect species abundances above a certain threshold (section 2.2). However, the relationship between the threshold at which a particular taxon is most likely unobservable and the immigration rate into the system is valuable when designing treatment plants.

A straightforward analytic approach is presented here. Assuming $N_T m(1-p_i)>1$ (which one might reasonably assume for real microbial systems), the combination of parameters is then sought for which there is no internal mode; in other words, when ϕ has no turning points in the range $[0,1]$.

Setting the time differential of ϕ in equation (3.23) to zero, we find the threshold is when

$$x_i\left(N_T m(1-p_i)-1\right)=(1-x_i)(N_T m p_i-1)$$

has no solutions in $[0,1]$. A little simple algebra reduces this condition to there being no solutions when $N_T m p_i \le 1$. Substituting this threshold value back into the calculation of the variance (equation (3.25)), it is seen that this threshold crudely corresponds to the variance of x_i rising above p_i^2. A similar calculation yields that the distribution is a skewed bell-curve when $N_T m(1-p_i)>1, N_T m p_i>1$.

3.2.2 Multidimensional Steady-State Distribution.

To find the equilibrium solution of the joint species distribution is a rather more complex task. The time differential in the multi-dimensional Fokker-Planck equation(3.17) is set to zero. Since the solution ϕ is a pdf, $\int \phi dx_1 ... dx_{n-1} = 1$ and by a similar argument to that employed in deriving equations (3.20)and (3.21), reflecting barriers are present at all the boundaries of the space $[0,1]^{(n-1)}$. Note that when there are n taxa in the community, this gives $(n-1)$ independent variables, as the sum of all n must necessarily be 1.

Therefore, $\forall i \in [1,2,...,(n-1)]$

$$0 = \left[-M_{\delta x_i} \phi + \frac{1}{2} \frac{\partial (V_{\delta x_i} \phi)}{\partial x_i} \right] + \frac{1}{2} \sum_{\substack{j=1 \\ i \neq j}}^{n} \frac{\partial (C_{\delta x_i \delta x_j} \phi)}{\partial x_j} \, \text{at} \, x_i = 0 \, \text{and} \, x_i = 1 \qquad (3.26)$$

Once more, ϕ is continuous and all boundaries of the domain $[0,1]^{(n-1)}$ can be regarded as being constant flux boundaries. This implies that, as with equation (3.22), the boundary conditions give that the integration constants are necessarily zero.

In other words, it suffices to find ϕ such that, for each $i = 1,...,(n-1)$

$$\frac{m(p_i - x_i)}{N_T} \phi = \frac{1}{2} \frac{\partial}{\partial x_i} \left(\frac{2x_i(1-x_i)}{{N_T}^2} \phi \right) + \frac{1}{2} \sum_{i \neq j} \frac{\partial}{\partial x_j} \left(\frac{-2x_i x_j}{{N_T}^2} \phi \right) \qquad (3.27)$$

Here it is shown that a Dirichlet distribution $Dir(N_T mp_1,\dots N_T mp_{n-1};N_T mp_n)$ satisfies these conditions. It should be noted, however, that the uniqueness of this solution has yet to be established.

Letting $X=(x_1,\dots,x_n)$, then if $X\sim Dir(N_T mp_1,\dots N_T mp_{n-1};N_T mp_n)$

$$\phi(X)=\Gamma(N_T m)\prod_{i=1}^{n}\frac{x_i^{N_T mp_i}}{\Gamma(N_T mp_i)} \tag{3.28}$$

Substituting the Dirichlet distribution (equation(3.28)) into equation (3.27) gives the following terms on the right hand side:

$$\frac{1}{2}\frac{\partial}{\partial x_i}\left(\frac{2x_i(1-x_i)}{N_T^{\,2}}\phi\right)=\left[\frac{\phi}{N_T^{\,2}}\right]\cdot \left[N_T mp_i-x_i(N_T mp_i+1)-\frac{x_i(1-x_i)}{x_n}(N_T mp_n-1)\right] \tag{3.29}$$

and

$$\begin{aligned}\sum_{i\neq j}\frac{\partial}{\partial x_j}\left(\frac{-x_i x_j}{N_T^{\,2}}\phi\right)&=-\frac{x_i\phi}{N_T^{\,2}}\sum_{i\neq j}\left[N_T mp_j-\frac{x_j}{x_n}(N_T mp_n-1)\right]\\&=-\frac{x_i\phi}{N_T^{\,2}}\left[N_T m(1-p_i-p_n)-\frac{1-x_i-x_n}{x_n}(N_T mp_n-1)\right]\\&=-\frac{\phi}{N_T^{\,2}}\left[N_T mx_i-x_i(N_T mp_i+1)-\frac{x_i\,(1-x_i)}{x_n}(N_T mp_n-1)\right]\end{aligned} \tag{3.30}$$

Adding these two quantities ((3.29)and (3.30)) gives

$$\frac{m(p_i - x_i)}{N_T}\phi$$

which is equal to the left hand side of (3.27), hence a $Dir(N_T mp_1,\ldots N_T mp_{n-1};N_T mp_n)$ joint pdf indeed solves the $(n-1)$ dimensional Fokker-Planck equation as required.

This result, which is that the joint probability density function for the abundance of all taxa in a neutrally assembled community is Dirichlet, is important for the practical implementation of neutral models. In the original formulation of the NCMs (Hubbell 2001), multiple realisations of the dynamics are required to generate the predicted community structure. This has been criticised by a number of authors, and analytic approaches offered (Vallade & Houchmandzadeh 2003; Volkov *et al.* 2003). Although sometimes mathematically very elegant, many of these are also idiosyncratic. For example, the solution offered by Volkov *et al.* (2003) is reliant upon the numerical integration across large factorial functions, which proved impossible for very large population sizes.

Knowledge that the abundance of all taxa can be generated from a single realisation of an appropriate Dirichlet distribution greatly simplifies the task of simulating communities, and sidesteps the need to employ some of the often cumbersome methods previously employed. What is more, the fact that we have a well known distribution opens up a great number of different established results.

3.2.3 Generating Realisations of Community Composition

Most standard mathematical and statistical packages (eg MATLAB) do not have a capacity to directly generate realisations of Dirichlet distributed random variables. However, an algorithm can be derived which allows for the realisation of Dirichlet distributed variables by normalising a series of Gamma distributed variables, which are more readily generated by such packages.

Let $y_1,\ldots,y_n$ be a set of n independent Gamma distributed random variables such that $y_1 \sim Gamma(N_T mp_1,1),\ldots,y_n \sim Gamma(N_T mp_n,1)$. Realisations of these variables can easily be generated by most standard mathematical and spreadsheet packages. It is shown here that if

$$x_i = \frac{y_i}{y_1 + \ldots + y_n} \qquad i = 1,2,\ldots,(n-1) \tag{3.31}$$

then the joint pdf for all x_is is given by the Dirichlet distribution in equation (3.28). Therefore to generate realisations of the abundance of each taxon in the community a realisation of each y_i is generated and then normalised as in equation (3.31).

To show that this is the case consider the pdf of each y_i

$$h(y_i) = \left[\frac{e^{-y_i} y_1^{N_T mp_i - 1}}{\Gamma(N_T mp_i)}\right] \qquad i = 1,\ldots,n\,.$$

Since the variables are independent, the joint density function for these n variables is simply the product of each of their pdfs, which is

$$f(y_1, y_2, \dots, y_n, y_r) = \left[\frac{e^{-y_1} y_1^{N_T m p_1 - 1}}{\Gamma(N_T m p_1)}\right] \cdots \left[\frac{e^{-y_n} y_n^{N_T m p_n - 1}}{\Gamma(N_T m p_n)}\right] \tag{3.32}$$

For $i = 1, 2, \dots, (n-1)$, let

$$x_i = \frac{y_i}{y_1 + \dots + y_n}$$

and

$$z = y_1 + \dots + y_n$$

Rearranging these gives

$$y_i = z x_i \tag{3.33}$$

and

$$y_n = z(1 - x_1 - \dots - x_{n-1}). \tag{3.34}$$

Substituting (3.33) and (3.34)into the joint density function(3.32) gives

$$g(x_1, \dots, x_n, z) = e^{-z} \left[\frac{(z x_1)^{N_T m p_1 - 1}}{\Gamma(N_T m p_1)}\right] \cdots \left[\frac{(z x_n)^{N_T m p_n - 1}}{\Gamma(N_T m p_n)}\right] \cdot |\det J|$$

where J is the Jacobian, given by $\begin{pmatrix} z & 0 & \cdots & 0 & x_1 \\ 0 & z & \ddots & \vdots & x_2 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & z & x_{n-1} \\ -z & -z & \cdots & -z & x_n \end{pmatrix}$

It can be seen that $|\det J| = z^{n-1}$, therefore

$$g(x_1,\dots,x_n,z)=\left[\frac{e^{-z}z^{N_Tm-1}}{\Gamma(N_Tm)}\right]\left[\frac{\Gamma(N_Tm)}{\Gamma(N_Tmp_1)\cdots\Gamma(N_Tmp_n)}\right]\cdot\left[x_1^{N_Tmp_1-1}\cdots x_n^{N_Tmp_n-1}\right].$$

The first term of this implies that $z \sim Gamma(N_Tm,1)$, and the second two brackets state that $x_1,\dots,x_n$ do indeed have a Dirichlet distribution with the required parameters. Additionally, $x_1,\dots,x_n$ are independent of z. Since Gamma variables are quickly and simply generated by standard mathematical packages such as MATLAB, such a technique is extremely easy to apply.

3.3 Frequency-p_i plots

The continuous neutral community model developed in the previous section is defined by the combined parameter pair N_Tm and the relative abundances of the taxa in the source community, $p_1,\dots,p_n$. As discussed in the literature review (section 2.3) there are two major assumptions made in Hubbell's original formulation; that the birth and death rates of all species are equivalent, and that the metacommunity from which immigrants are drawn has a logseries taxa abundance distribution (i.e. the p_is are distributed according to a logseries).

This assumption on the logseries nature of the source community dictates that it can be characterised by a single parameter, θ, the fundamental biodiversity number. In all previously published applications of NCMs, irrespective of the precise mathematical formulation of the model, N_T has been estimated directly from survey data. m and θ have then been estimated by minimising the differences between estimates and observed taxa-abundance distributions for the whole community (Hubbell 2001; Dunbar *et al.* 2002; Volkov *et al.* 2003).

For microbial communities, the number of individuals in a sample can be estimated by a direct count and, by assuming a constant density of microorganisms, the community size, N_T, can be calculated. Whatever the method employed, whether clone libraries, DGGE or some other molecular method, the laboratory procedures are somewhat laborious and, as previously discussed (section 2.2), cannot begin to offer a complete picture of the entire community. Accordingly, the m and θ parameters cannot be estimated for microbes in the same way as for larger organisms. Without the luxury of full taxa-abundance distributions upon which classical ecologists can rely, a novel approach is required for microbial datasets. Owing to both detection limitations of some techniques such as DGGE, and the unavoidable sampling issues previously discussed, a calibration procedure is required which is reliant solely upon the data which can be gathered from the observable tail of the true taxa-abundance distribution.

This task of calibrating parameters from scarce sample data can be made simpler and more attainable for microbial datasets by uncoupling the two central assumptions in Hubbell's formulation of his Neutral Community Model. In other words, a NCM can be defined which assumes the same neutral community assembly and dynamics in a local community as Hubbell's NCMs, but without the additional assumption of a logseries distributed source community for immigration. In this section, a method is proposed for calibrating a neutral model that requires no additional assumption about the nature of the metacommunity. That is, the $p_1,\dots,p_n$ values do not have to be realisations of a logseries distribution, although they may be. Where suitable data are available from a number of similar but distinct communities, the distribution of the abundances of the common taxa in the metacommunity can be estimated from the information gathered, rather than by assuming some underlying evolutionary process, which we would be uncertain of for microbes.

Given survey data, the task facing mathematical modellers is to extrapolate information about these. A simple count or a reliable estimate of N_T is a simple enough procedure, but the rate of immigration is not something which can be straightforwardly measured. However, for datasets consisting of taxa abundance information at a number of similar sites, a simple technique is presented here for estimating the combined $N_T m$ parameter as well as the metacommunity abundances, $p_1,\dots,p_n$ under the assumption that all such sites can be described by realisations of the same abundance distribution.

Since the NCM predicts that the taxa abundances will have a joint distribution which is Dirichlet distributed $Dir(N_T mp_1,\dots N_T mp_{n-1};N_T mp_n)$, estimators of the relative abundances in the source community can be derived empirically. The expected abundance of species i at each measured site is $E(x_i)=p_i$ for $i=1,\dots,n$. Therefore, if data are available from a sufficient number of similar sites, the mean relative abundance of each taxon over all local sites is equal to the relative abundance of that taxon in the metacommunity. Accordingly, each p_i value is estimated by the mean abundance of taxon i across all communities in the dataset. The advantage of this method of characterising the source community is that it sidesteps the need for a conceptual picture of the evolution and dynamics of metacommunities, as in Hubbell's original work, and also eliminates one of the parameters, θ.

Thus p_i and N_T can be measured; p_i from enumerating specific species at many similar sites and N_T from counting microorganisms. Therefore, the only model parameter that remains to be estimated is the immigration probability m.

The frequency-p_i method centres on plotting a curve of the estimated p_i values for each taxon detected against the number of sites at which it is found. In general, and perhaps unsurprisingly, the more abundant taxa in datasets tend to be far more ubiquitous than the rarest ones, which tend to appear in samples from only a small proportion of the locations (see figure 3.2). This trend seems to support the idea of there being some stochastic element to community assembly.

The relationship between p_i and frequency can be formalised for the neutral model. The knowledge that all the marginal densities of species $x_i \sim Beta\left(N_T m p_i, N_T m(1-p_i)\right)$ allows for calculation of the probabilities that a given taxon appears present at a given site in the dataset. In order for each taxon to be deemed present at a location, there must be at least one member of that species at the site. That is, for a community of size N_T, the abundance of taxon i must satisfy $x_i \geq 1/N_T$ for it not be appear absent at that location.

Then, under the assumptions of the NCM, it is the case that the probability of a given species being observed at each site is

$$P(x_i \geq 1/N_T) = \int_{1/N_T}^{1} \phi(x_i; p_i, N_T m) dx_i \tag{3.35}$$

where $\phi(x_i; p_i, N_T, m)$ is the marginal probability density function of the local abundance of a given taxon, which has previously been shown to have a Beta distribution $Beta[N_T m p_i, N_T m(1 - p_i)]$. This procedure is applied for all taxa present at any site in the whole dataset. Strictly speaking, the presence/absence of each taxon is not independent of all others (as is implicitly assumed here by integrating each marginal density separately rather than a multi-dimensional integral of the joint density function). However, given the immense population sizes of real microbial systems, these conditional probabilities prove close enough to the individual probabilities of considering each taxon individually for the differences to be considered negligible. Additionally, at this stage of the thesis, the detection limitations and biases inherent to molecular methods (section 2.2) are neglected. These issues are more fully examined in the following chapter.

Just as the abundances of all taxa were averaged to estimate the metacommunity abundances, it is taken that proportion of sites at which each taxon is present closely approximates the probability of it being detected at a randomly selected location.

A plot of the metacommunity abundance of each taxon against the number of sites at which it is detected is then produced. Then, for a given value of the $N_T m$ parameter, a curve can be fitted through these points, which represents the frequency-p_i curve as predicted by the NCM. By varying the $N_T m$ parameter, the errors between the prediction and the data points can be minimised. For the curves presented here (figure 3.2), the error minimisation techniques was a simple least squares fit. More such parameterisations of the model are presented in Sloan *et al.* (2006)

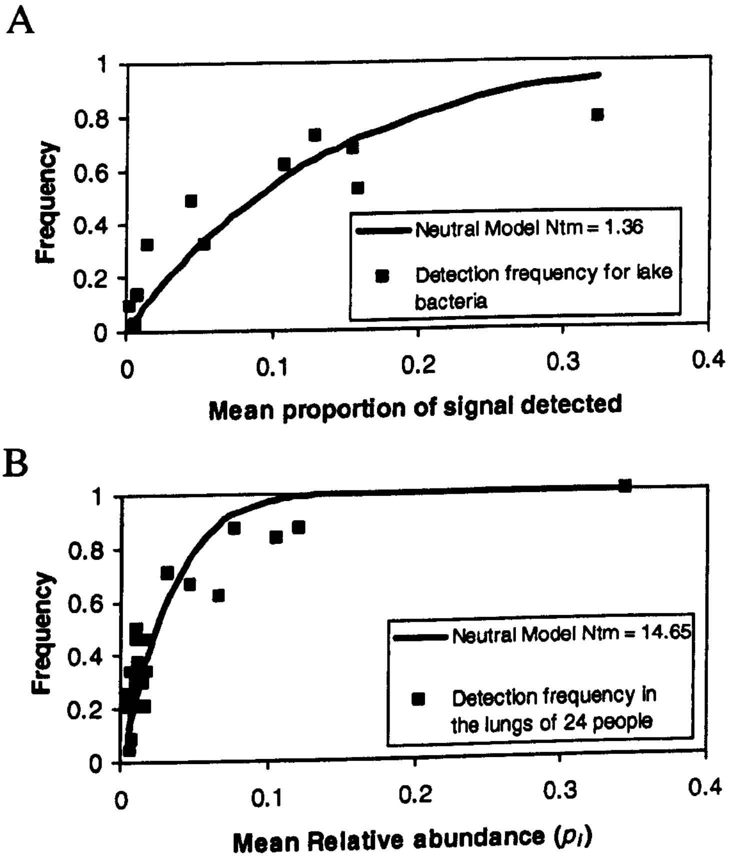


Figure 3.2

Comparing the theoretical and observed relationship between the mean relative abundance of a taxon, p_i, and the frequency with which it appears in a fixed population size. Each of the points represents a different taxon. (a) 16S RNA sequences for 16 different bacterial taxa that are considered to be particular to freshwater environments sampled from 96 different lakes(Zwart *et al.* 2002). (b) Clones from the lungs of 24 people with and without asthma. Sloan *et al.* (2006)

These frequency-p_i plots provide a preliminary test as to whether a dataset could have been collected from a neutrally assembled community. If there were no limitation on the dispersal of microbes, then each local community could be regarded merely as being a sample from some larger community. If this were the case, then the calibrated immigration rate would always be 1 and the stochastic effects of random sampling would mean that taxa were absent from some communities and present in others purely by chance. For example, if the relative abundance of an organism in the source community is 0.5 then it is very likely to be seen in samples from all locations, but an organism with relative abundance 0.001 will rarely appear in small communities. This then begs the question, how much of an effect does immigration have on the community structure over-and-above that which is due to random sampling from identical communities? If the local communities were unbiased random samples from some identical source community then p_i, the mean relative abundance, is the probability that an organism picked at random from the local communities belongs to the i^{th} taxon. Thus if K is the number of individuals in a community of size N_T that belong to the i^{th} taxon, it will be distributed binomially,

$$P(k = K) = \binom{N_T}{k} p_i^{\,k} (1 - p_i)^{N_T - k}.$$

Fig 3.3 below illustrates this. In the first frequency-p_i plot, the observed data fit the expected shape for a neutrally assembled community, but are clearly not binomially distributed, thus providing evidence of the role dispersal limitation plays in shaping community structure. In the second plot, the shape of the frequency-p_i curve does not differ significantly from that of a binomial distribution. It therefore cannot be ruled out that these organisms represent samples from the same common source community, and thus this dataset does not offer any evidence to support NCMs. There is no evidence of dispersal limitation being the driving force behind community structure; immigration alone suffices as an explanation.

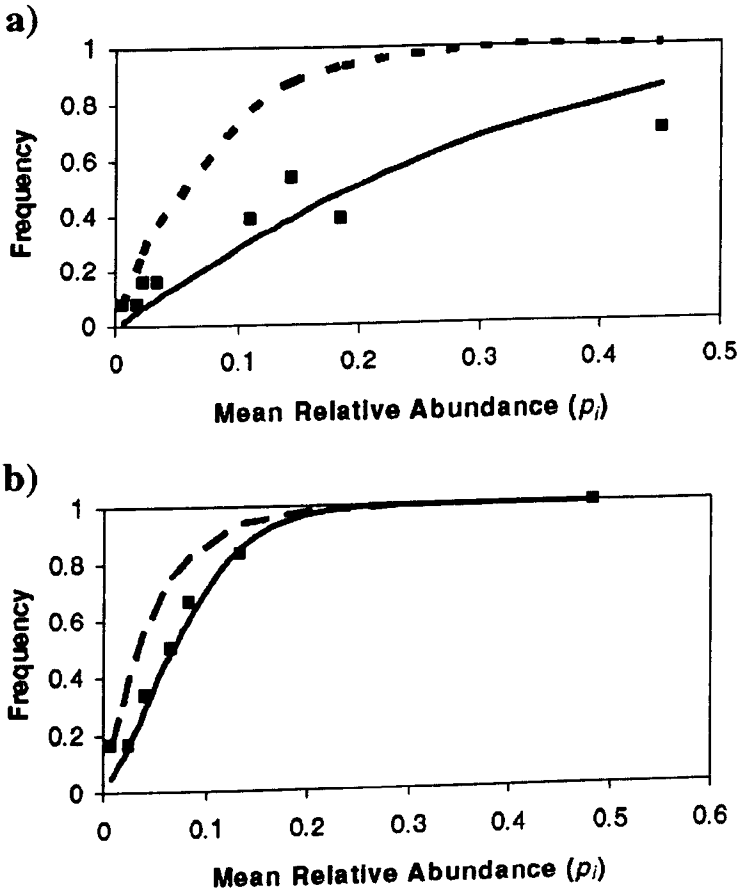


Figure 3.3
Comparing the theoretical and observed relationship between the mean relative abundance of a taxon, p_i, and the frequency with which it appears in a fixed population size. The solid lines represent the best fit to a NCM and the dashed lines are assuming that the taxa abundances are binomially distributed. a) AMO genes at 13 different domestic sewage works (Wagner & Loy 2002). b) AOB genes at six sites from the Humber Estuary (Linacre 2004).

Furthermore, it is not simply the case that almost any community can appear to be either neutrally or else purely randomly assembled from some metacommunity. Fig 3.4 below shows a dataset for bacterial communities collected from samples of human faeces. This is seen to provide a much less convincing fit to a NCM. In this case, it appears that subtle genetic differences between the bowels of the various test cases are a far stronger factor in shaping the communities than are the simple mechanisms inherent to the neutral models.

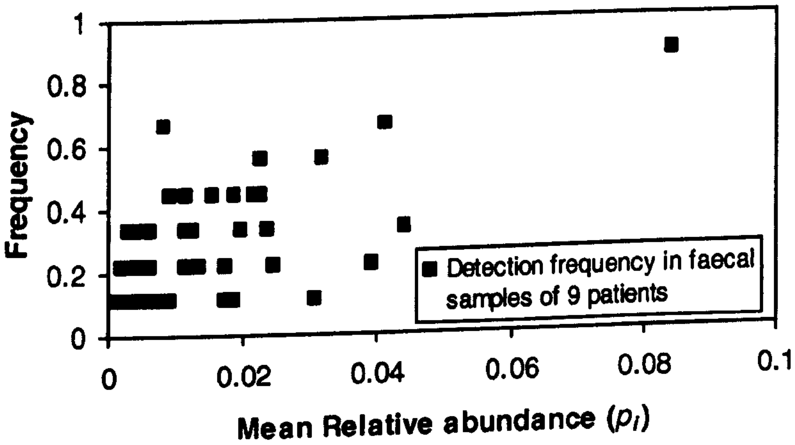


Figure 3.4
Comparing the theoretical and observed relationship between the mean relative abundance of a taxon, p_i, and the frequency with which it appears in a fixed population size. The dataset here is for clone libraries of microbes detected in faecal samples from 9 people taken from Mangin *et al.* (2004). Note that the pattern does not appear to support the idea of neutral community assembly.

3.4 Conclusions

Neutral Community Models have been shown to be successful in many different communities. It was argued in this chapter that previous published versions of NCMs were inappropriate for application to microbial communities for two reasons; the models were discrete and parameterisation required vastly more data than current molecular methods could hope to provide. These two limitations of previous versions of NCMs have been overcome in this chapter. The NCM as presented here is a continuous model capable of predicting communities of arbitrarily large size. Additionally, a new calibration procedure was developed which can be used with currently obtainable data. It has been demonstrated that this continuous NCM can be calibrated to fit existing datasets. Calibration, however, is not validation of a model. Whilst this chapter provides evidence in support of NCMs, many other models could exist which produce similar patterns. For example, McGill *et al.* (2006) argue that niche differentiation models better explain the biological complexity which is known to exist.

At this stage of the thesis, these criticisms have not been addressed, and it is accepted that other models could explain the data. However, the comparison with the binomial model does lend further evidence to support the hypothesis that exactly the same structuring forces do not act on all environments and that the effects of dispersal limitation can be seen in multiple samples from similar environments. Later in the thesis, a remarkable dataset is presented that provides much stronger evidence in support of neutral community assembly.

The research in this chapter demonstrates that a stochastic NCM can be used to describe the patterns of community structure in the microbial world at the scale at which they are typically observed and using measurement methods that are routinely employed. Indeed it is possible that NCMs will find their widest application in microbiology since a number of the assumptions inherent in neutral models may typify much of the microbial world.

4 Sampling Effects in Microbial Modelling

In the previous chapter of this thesis, the Neutral Community Model developed by Hubbell for macroorganisms was refined and a similar model developed for describing the assembly and structure of microbial systems. NCMs in classical ecology have been shown to reproduce the fundamental patterns in nature which ecologists have been trying to explain for decades (Hubbell 2001a; Volkov *et al.* 2003; Sloan *et al.* 2006). The central premise of this thesis is that similar models in microbial ecology will prove to be equally fruitful forms the central premise of this research.

There is, however, a generic problem in applying any model to microbial communities, whether a Neutral Model or one of the alternatives outlined in the literature review in Chapter 2 (for example the Lognormal), the problem faced is the same. Even using the most up to date laboratory techniques, it is only possibly to observe the structure of a tiny subsample of what are ostensibly large communities. Therefore, it is necessary to infer patterns that might occur in the larger system.

Any mathematical model of microbial community structure, if it is to be applied to make predictions on real-world systems, requires calibration using data gathered from other similar communities. It is therefore imperative that any theoretical models applied to microbial systems are developed with a keen eye on the laboratory techniques and the sorts of sample data that they can provide; an elaborate mathematical model that requires unobtainable information for parameterisation is of little to no practical use.

For studies in classical ecology, consideration of sampling effects is an established and much discussed problem, one for which a number of techniques have been developed to quantify and correct any biases which arise (Colwell & Coddington 1994; Chao *et al.* 2005). The scale of the sampling issues faced by microbial ecologists, however, is much greater than those in classical ecology and these techniques are not always applicable. Given sufficient time and resources, a complete census of all the trees in a forest is certainly possible. An exhaustive count of any microbial system larger than that on the head of a pin is a rather less attainable goal. It is argued here that the magnitude of the unavoidable undersampling of microbial communities is almost always immense. The discrepancy between the sizes of microbial communities and the number of individuals whose DNA can be characterised in samples from the communities typically spans many orders of magnitude. Take for example a study of soil based microbes. In each gram of soil, there are of the order of 10^{10} prokaryotes (Whitman et al. 1998). Current laboratory tools cannot hope to take an exhaustive sample from such a population. As discussed in chapter 2, if 16s rRNA clone libraries are employed, sample sizes are of the order of tens to hundreds. With DGGE analysis, sample sizes may be of the order of around a million individuals, but even this represents significant undersampling. By analogy, when there are currently around 6×10^9 humans in the world, a single sample of a few hundred individuals is unlikely to be sufficient to characterise the global distribution of any human trait unless it is extremely homogeneous.

In this chapter, the biases that current laboratory tools and techniques introduce to classical measures of community composition are investigated and a method for compensating for them when characterising a neutrally assembled community is developed.

The main achievements in this chapter are:

- Demonstration of the need to account for sampling effects in characterising microbial communities. Although a number of correction techniques exist for sampling biases for studies in classical ecology, where representative samples can be collected, none have been developed specifically for the problem of undersampling on the scale of that faced by microbial ecologists. Here, a similar technique is presented for extrapolating information from samples as small as those currently analysed by modern molecular methods.
- Further development of the Neutral Community Model such that in addition to predicting the community structure within a given system, it also predicts the distribution of taxa-abundances in small samples taken from it. This helps to provide quantifiable estimates of what will be observed using current microbial tools.
- Refinement of the frequency-p_i plot technique to consider the effects of sampling upon presence/absence datasets. For many of the cases where the sampling procedures employed distort the view of the immigration rate, a simple extrapolation rule is developed to give an estimate of the true immigration into the whole system.

These achievements form the basis of publications in Microbial Ecology (Sloan *et al.* 2007) and Philosophical Transactions of the Royal Society of London (Curtis *et al.* 2006).

4.1 Modelling The Sampling Effects Typically Encountered When Employing Current Microbial Tools

The premise of this thesis is that the wealth of new laboratory tools being developed for observing and characterising microbial systems has the potential, when combined with theory, to revolutionise the design of engineered biological systems. That said, the enthusiasm for such emerging technologies should not be unrestrained. These exciting new tools currently offer only partial and fleeting glimpses into the microbial world and cannot begin to provide the complete surveys that are often achieved for communities of macroorganisms. For example, taking a census of trees in a forest simply requires the correct identification and numeration of each tree species, a lengthy but certainly possible task. Even the most advanced tools for microbial community analysis are complex, multi-stage operations which can only ever examine a tiny fraction of the sorts of populations observed in real life engineered systems.

Take for example a ten gram sample of soil; this can comprise as many as 10^{10} individual microorganisms. Clone libraries generated from soil samples typically represent a random sample of tens to a couple of hundred individuals. Intuitively such small samples have the potential to distort the view of the large population. Whilst microbial ecologists are well aware of this disparity of scale, it does not routinely affect the way laboratory data are interpreted. Taxa abundance distributions, for example, are used to characterise microbial community structure but what they actually characterise is the distribution of taxa abundances in a very small sample. The disparity of scale between samples and the communities they aim to represent, mean that the sample and community distributions can be very different indeed.

This point is graphically demonstrated in figure 4.1. The taxa abundance distributions for four large populations (each of size 10^{12} individuals) with different taxa abundance distributions are plotted alongside the typical distributions observed in samples. In each case, the sample distribution is generated for 200 selected at random from the large populations. This is equivalent to the number of clones in a typical 16S rRNA gene clone library, assuming no additional biases are introduced during the PCR amplification stage (section 2.2). This is done for four communities in which the taxa abundances are distributed in four very different ways (Figure 4.1); two of which have been previously been proposed as plausible theoretical distributions (Finlay & Clarke 1999; Curtis *et al.* 2002a) (Figure 4.1a, Figure 4.1b) and two of which have no biological basis and could be considered ridiculous (Figure 4.1c, Figure 4.1d). All the sample distributions have a very similar shape which is redolent of the distribution of taxa-abundances in real 16S rRNA gene clone libraries (Wagner 2000). Thus, for example, the fact that clone abundance distributions look like the tail-end of a lognormal distribution does not mean that the taxa in the larger community are distributed lognormally (although they might be). Descriptors such as diversity indices, taxa-abundance distributions and similarity indices have their roots in ecology of macroorganisms, which are easier to observe, and rely on a fairly complete census of the organisms at a particular site. Figure 4.1 demonstrates that for microbial communities these descriptors may differ significantly between sample and community. Molecular methods are rapidly evolving and will offer partial solutions. Thus when very high throughput sequencing becomes routinely available to microbial ecologist a complete census of a sample may become possible. However, improved molecular methods will offer only part of the solution; the number of individuals in samples, even when they can all be identified, will still be very small in comparison to those in the microbial communities as a whole. It will always be necessary to infer larger-scale descriptors of community structure from very small samples, which requires consideration of sampling effects. Nonetheless, the fact that patterns exist in the common taxa suggests generic patterns that might extend deeper into the community.

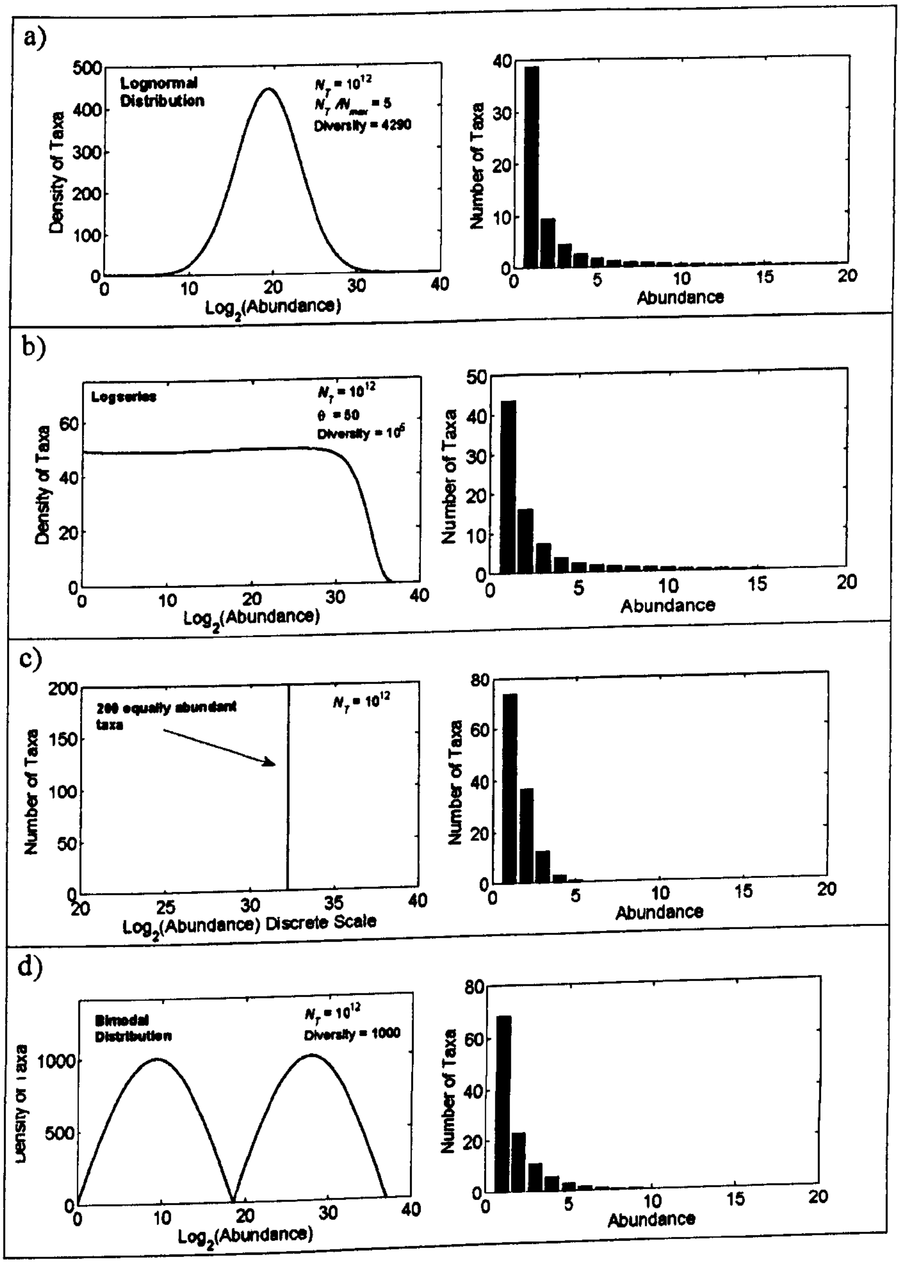


Figure 4.1

Distribution of taxa abundances in communities of 10^{12} individuals and in small samples of 200 from them for a) a lognormally distributed community, $N_T / N_{\max}$ is the ratio of the total number of individuals to the number of individuals belonging to the most abundant taxon, which can be used to index diversity (Curtis *et al.* 2002b). b) a logseries distributed community, θ is the parameter used to index diversity (Hubbell 2001a). c) a community where 200 taxa are equally abundant. d) a bimodal distribution.

4.2 Sampling from a Neutral Community Model

It has already been shown in section 3.2 that for the continuous variant of the Neutral Community Model (NCM), the steady state joint probability density function for all species is Dirichlet $Dir(N_T mp_1,...N_T mp_n)$ where $p_1,...,p_n$ are the relative abundances of the taxa in the metacommunity. However, as previously demonstrated, this distribution can be somewhat different to that observed in small samples.

Here, an analytic form of the approximate joint density function for all species within a sample of N_S individuals from a larger neutrally assembled community is derived. A similar argument is employed as was in Chapter 3 for finding the local community abundance distribution.

Strictly speaking, selecting a subsample of size N_S from a local community is achieved by sampling N_S individuals without replacement from the community of size N_T. Here, the problem is approximated to one of sampling with replacement, since for almost all microbial samples $N_S \ll N_T$, so that the chance of any given organism being sampled twice from the local population is negligible.

The sampling exercise can then be considered as a continuous process through time. Individuals are selected from the source community one by one until a sample of size N_S is obtained. Once this sample size has been reached, the process of selecting individuals is continued at regular intervals in time (generations) but now the selected individual replaces one randomly chosen individual currently in the sample population. This is analogous to the argument used for deriving the joint distribution for the local abundances, except that it is a pure immigration-death process, with immigrants into the sample from the local community. Setting $m=1$ and regarding the local neutrally assembled community as the source from which immigrants are drawn then, by the results in Chapter 3, conditional on knowledge of local community abundances $x_1,\ldots,x_n$ the joint distribution of relative abundances $y_1,\ldots,y_n$ within a sample is Dirichlet $Dir(N_S x_1,\ldots N_S x_n)$. That is,

$$f(Y \mid X) = \Gamma(N_S) \prod_{i=1}^{n} \frac{y_i^{N_S x_i}}{\Gamma(N_S x_i)} \tag{4.1}$$

where $X = (x_1,\ldots,x_n)$ and $Y = (y_1,\ldots,y_n)$ for notational convenience.

Now, both the unconditional joint density function of the taxa abundances within the local community ($X \sim Dir(N_T m p_1,\ldots,N_T m p_{n-1}; N_T m p_n)$ and the conditional joint density function of the sample abundances ($Y \mid X \sim Dir(N_S x_1,\ldots,N_S x_{n-1}; N_S x_n)$) are known. By conditional probability theory, the unconditional joint distribution of the sample abundances can be obtained directly by integrating across all the values the values the (n-1) dimensional variable X could take.

In other words, the joint distribution for the relative abundances within a sample from a NCM is known to be

$$f(Y) = \Gamma(N_S) \cdot \Gamma(N_T m) \int_{[0,1]^n} \prod_{i=1}^{n} \left[\frac{y_i^{N_S x_i}}{\Gamma(N_S x_i)} \cdot \frac{x_i^{N_T m p_i}}{\Gamma(N_T m p_i)} \right] dX \tag{4.2}$$

Unfortunately, evaluation of this integral seems to be a non-trivial task and. it was not possible to find an analytic solution to this equation. Accordingly, an alternative approach to the problem was pursued.

In the absence of a neat analytic solution of equation(4.2), multiple realisations of sampling from NCMs were generated and the probability density functions for the abundance of each taxon then approximated. Visual examinations of these approximate sample distributions for each individual taxon seemed to suggest that each of the sample abundances, $y_1,\ldots,y_n$ is also Beta distributed with the same expectations as in the local community, but with different variances.

Accordingly, the approach adopted was to seek to fit a Dirichlet distribution (which has Beta marginal densities) with parameters chosen to ensure that the first and second moments match those known for the sample distributions. These are, respectively

$$E(y_i \mid x_i) = x_i \tag{4.3}$$

and

$$E(y_i^2 \mid x_i) = \frac{x_i(N_S x_i + 1)}{(N_S + 1)} \tag{4.4}$$

Now, since $x_i \sim Beta(N_T m p_i, N_T m(1 - p_i))$ we have that

$$E(x_i) = p_i \tag{4.5}$$

and

$$E(x_i^2) = \frac{p_i(N_T m p_i + 1)}{(N_T m + 1)} \quad (4.6)$$

Then, by the elementary laws of conditional expectations,

$$E_{Y_i}(y_i) = E_{X_i}\left[E_{Y_i|X_i}(y_i \mid x_i)\right] \quad (4.7)$$

By substituting in from equation (4.3)

$$E_{Y_i}(y_i) = E_{X_i}\left[x_i\right].$$

Then, equation (4.5) gives

$$E_{Y_i}(y_i) = p_i \cdot \quad (4.8)$$

Applying the same technique to the second moment gives

$$E_{Y_i}(y_i^2) = E_{X_i}\left[E_{Y_i|X_i}(y_i^2 \mid x_i)\right] \cdot \quad (4.9)$$

Applying equation (4.4) gives

$$E_{Y_i}(y_i^2) = E_{X_i}\left[\frac{x_i(N_S x_i + 1)}{(N_S + 1)}\right] \quad (4.10)$$

and substituting in the results (4.5) and (4.6) gives

$$E_{Y_i}(y_i^2) = \left[\frac{1}{N_S + 1}\right]\left[N_S \frac{p_i(N_T m p_i + 1)}{N_T m + 1} + p_i\right] \quad (4.11)$$

$$= \left[\frac{1}{N_S + 1}\right]\left[N_S \frac{p_i(N_T m p_i + 1)}{N_T m + 1} + p_i\right]$$

$$= \frac{N_S N_T m p_i^2 + (N_T m + N_S + 1) p_i}{N_S N_T m + N_T m + N_S + 1}$$

$$= \frac{\left(N_S \frac{N_T m}{N_T m + N_S + 1}\right) p_i^2 + p_i}{\left(N_S \frac{N_T m}{N_T m + N_S + 1}\right) + 1} \quad (4.12).$$

It is now observed that, defining

$$\tilde{m} = \frac{N_T m}{N_T m + N_S + 1} \tag{4.13}$$

gives

$$E(y_i) = p_i$$

and

$$E(y_i^2) = \frac{(p_i N_S \tilde{m} + 1)p_i}{N_S \tilde{m} + 1} \tag{4.14}$$

implying that, correct to the first two moments, the distribution of y_i, the sample abundance, can be taken to be approximately Beta distributed $y_i \sim Beta(N_S \tilde{m} p_i, N_S \tilde{m}(1 - p_i))$ where $\tilde{m}$ is as defined in(4.13).

The main consequence of this result is that, when working with sample data, direct measurement of the community immigration rate, m, is not possible. Instead, what is observed is an 'effective immigration rate', $\tilde{m}$, into samples. This phenomenon is easily overlooked, and it is imperative that estimates of the immigration rate derived from sample data acknowledge the relationship between the actual immigration rate, m, and the observed sample rate

$$\tilde{m} = \frac{N_T m}{N_T m + N_S + 1}. \tag{4.15}$$

The significance of this is that, once the immigration rate using samples has been calibrated (for example by the frequency-p_i method), the immigration rate for the entire community must then be extrapolated from this information by the above relationship. The following section discusses the implication of this result upon the frequency-p_i method of data fitting.

4.3 Sampling Effects upon the Frequency-p_i calibration method for NCMs

In section 3.3, a simple method was proposed for calibrating the key immigration parameter, m, in Neutral Community Models. Using a multi-location dataset, the abundance of each taxon in the metacommunity, p_i can be estimated simply by taking its mean relative abundance across all observed communities. The frequency of each taxon's detection is then taken as the proportion of sites in the dataset at which it was observed. Finally, a plot is produced of the p_i values against the frequencies of detection. A NCM can then be fitted with $N_T m$ (and implicitly therefore, m) chosen such that the squared errors between the model's predicted frequencies and those observed in the dataset are minimised.

If sampling effects are to be fully considered, there are two additional issues which arise, one concerning the estimation of the detection frequencies, the other the fitted immigration parameter.

In order to address the first of these problems, one major question must be answered; how do presence/absence datasets for samples relate to presence/absence datasets for complete communities? Surely, if a particular taxon is detected present in any sample, it is also present in the system sampled. The converse, however, is not the case. Even if detection of species within a sample is assumed to be complete, there is still a sizeable chance that a particular taxon present in the system is simply not selected during the sampling procedure and so appears absent when it is indeed in the larger community. With the massive undersampling inherent in microbial analysis, this cannot be overlooked. There is also the additional issue of laboratory biases and detection limitations.

In section 2.2, the most commonly employed microbial tools employed in laboratory work are discussed, along with some of their limitations. Here, the relative merits of Denaturing Gradient Gel Electrophoresis (DGGE) and of Clone Libraries are considered. The most obvious difference between these two different tools is in terms of throughput of DNA; with DGGE, samples of the order of a million or so microbes can be analysed whereas Clone Libraries typically consist of only hundreds or even tens of microorganisms. The major advantage of the later technique over DGGE, however, is that there are no detection limitations and so there is a finite probability of observing every individual in the community. A single member of a taxon in a Clone Library will appear present, whereas a singleton will not produce a band on a DGGE gel sufficiently bright for detection. Typically, abundances must be of the order of somewhere between 0.1% and 1% of the total biomass for a taxon to be detected on a gel (Cocolin *et al.* 2000; Woodcock *et al.* 2006).

These detection limits have a knock-on effect on the calculation of the predicted detection frequencies with NCMs. When working from sample data, presence/absence information does not reflect the presence or absence of each taxon in the larger community. Indeed, when there is a detection threshold, as with DGGE analysis, presence/absence datasets do not even provide this information for the sample itself. All that can be ascertained is that each taxon will only appear if its abundance within the sample is greater than the detection limit.

Such complications do not, however, prevent the application of the frequency-p_i calibration method to sample data. The knowledge of each laboratory tool is sufficient, however, to allow estimates of d, the detection limit such that if $p_i \geq d$ then a taxon is observed, else it appears absent. For Clone Libraries, a singleton in a sample of size N_S will be detected, so $d = 1/N_S$. If DGGE analysis is employed then typically $0.001 < d < 0.01$ (Cocolin *et al.* 2000).

Then, under the assumptions of the NCM, it is the case that the probability of a given species being observed at each site is

$$P(x_i > d) = \int_d^1 \phi(x_i; p_i, N_T m) dx_i \tag{4.16}$$

where $\phi(x_i; p_i, N_T, m)$ is the marginal probability density function of the local abundance of a given taxon, which has previously been shown to have a Beta distribution $Beta[N_T m p_i, N_T m(1-p_i)]$.

There is the additional issue of biases which may well creep in during the PCR amplification stage associated with both techniques. A study of such effects is beyond the scope of this thesis. However, there are promising, but as yet not fully explored, theoretical routines to quantifying and correcting these biases(Jagers & Klebaner 2003).

The second issue is that of extrapolating information about immigration into the whole system from sample data. As shown in the previous section of this chapter, the immigration rate observed into samples is not the same as that into the whole community itself. The frequency-p_i plot method can then be applied using the correct detection limit and the observed immigration rate calibrated. However, equation (4.15) shows that the observed effective immigration (as defined in section 4.2) into samples, $\tilde{m}$,is related to the sample size, N_S,and the $N_T m$ parameter by

$$\tilde{m} = \frac{N_T m}{N_T m + N_S + 1}.$$

This can, however, create a problem in certain cases. Initially, for small values of $N_T m$ (relative to N_S) the changes in the effective immigration rate, $\tilde{m}$ are first order in powers of $N_T m$. However, as this combined $N_T m$ parameter increases relative to a constant sample size, the rate at which changes in $\tilde{m}$ are observed decreases, with $\tilde{m}$ finally tending towards unity.

Obviously, this issue is particularly troubling when attempting to fit the model to datasets of samples from high immigration communities. For more stagnant systems, small errors in calibrating the effective immigration into the sample will result in similarly small errors in the extrapolated parameter $N_T m$. However, when the immigration rate is much higher ($N_T m \gg N_S$) $\tilde{m}$ approaches unity and it becomes extremely difficult to determine the $N_T m$ parameter with any degree of accuracy at all. This problem is especially pronounced when handling clone library datasets, for which N_S is typically very small.

For example, consider the examination of a library of around 100 clones. If the effective immigration rate is determined within 10% error margins to be around 0.90, then $N_T m$ is extrapolated using equation (4.15) and a value of 910 is calculated. However, if the calibrated $\tilde{m}$ value should correctly have been either 0.81 or 0.99 (both within the hypothetical error limits) then the true $N_T m$ value could be as little as 430 or as great as 10000. The difference between these two very dissimilar communities is virtually undetectable in small clone libraries.

No such problems arise, however, for systems which are subject to much lower immigration rates. As before, consider a library of 100 clones and assume the effective immigration rate is calculated within 10% error margins to be around 0.02. Extrapolation of the true $N_T m$ value for this case suggests that it lies between 1.85 and 2.27.

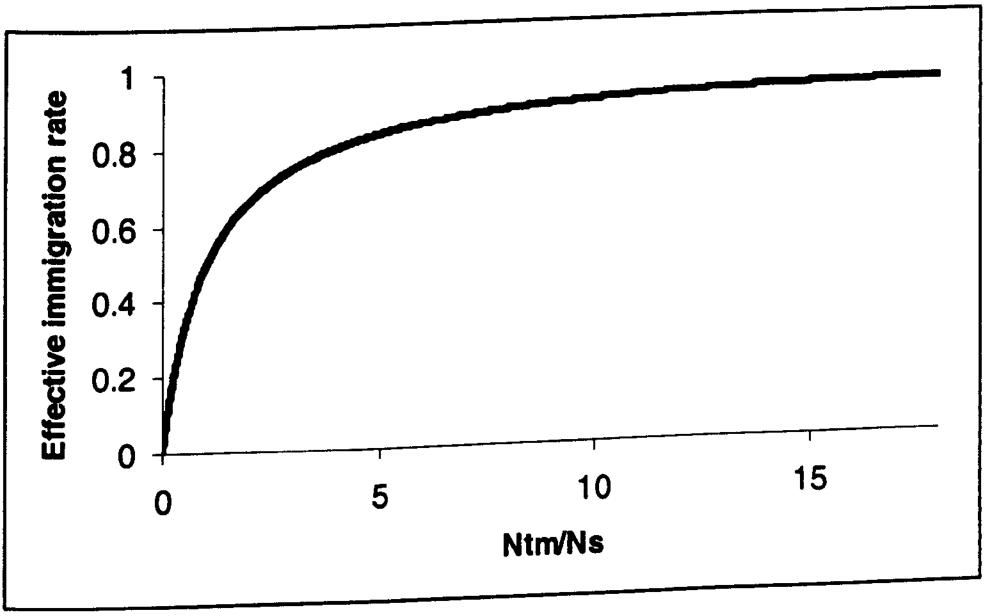


Figure 4.2

Effective immigration rate into samples plotted as a function of the ratio $N_T m / N_S$. It is assumed that $N_T m + N_S \gg 1$.

This is perhaps best explained by considering two of the example data sets displayed in Figure 4.3: clone libraries of ammonia monooxygenase AMO genes(Purkhold *et al.* 2000; Wagner & Loy 2002a) from 13 different sewage works in Germany; and the ammonia oxidising bacteria 16S rRNA gene data from 6 samples at three different sites in the Humber estuary in England (Linacre 2004b). On average 13 clones were sampled from each of the sewage work samples and exactly 20 were sampled for the estuary samples. As argued previously this is a small sample from which to draw conclusions on the community structure at any one site. However, using the frequency-p_i fitting method, it is possible to calibrate the neutral community model based on the distribution of taxa-abundance across the 13 sewage works or 6 estuary samples for the common taxa.

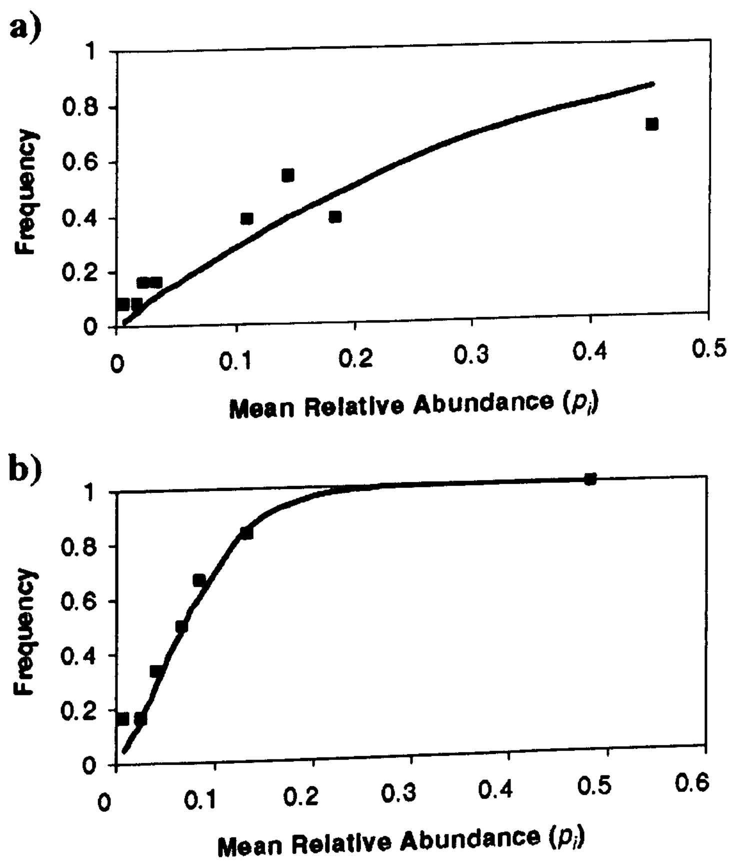


Figure 4.3

Comparing the theoretical and observed relationship between the mean relative abundance of a taxon, p_i, and the frequency with which it appears in a fixed population size. a) AMO genes at 13 different domestic sewage works(Wagner & Loy 2002b). b) AOB genes at six sites from the Humber Estuary(Linacre 2004a). The best fit parameters are a) $\tilde{m} = 0.1$ and b) $\tilde{m} = 0.77$.

In both studies, the immigration rate can be simply calibrated by adjusting it to minimise the difference between this theoretical probability of detection and the observed relative frequency with which the common taxa are observed. For the sewage works samples the calibrated rate of immigration is 0.1 and for the estuary samples it is 0.77; this is the probability that when an ammonia oxidising bacterium is lost from the system it is replaced from outside.

At first glance, this may seem to suggest that immigration of AOB into German sewage works and into samples from the Humber estuary is high and perhaps dispersal limitation is not a major driver in shaping community structure in these communities. However, once the sampling procedures are taken into account and the extrapolation factor from sample scale to community scale is employed (equation (4.15)) a very different result is found.

The calibrated value of immigration from the samples taken is in fact $\tilde{m}$, the effective immigration rate into the small sample that encapsulates both the dispersal limitation imposed on the community as a whole and random sampling effects. Equation (4.15) allows for the extrapolation from small random samples to the immigration in the larger neutral community. In the case of the sewage works, where the effective immigration probability is 0.1, the immigration probability for a neutral community of 10^9 organisms would be 1.55×10^{-9}. For the estuary, where the effective immigration is 0.77, the immigration probability for a neutral community of 10^9 organisms is an order of magnitude higher, but still low at 7×10^{-8}. This would indicate that immigration for both environments is low if a representative element of the microbial landscape comprises 10^9 organisms.

As previously discussed (fig 4.2), it is apparent that for the effective immigration into a sample to vary significantly from 1 then the product $N_T m$ must be at most of the order N_S. This means that when the sample size, N_S, is small and N_T is large, the immigration probability has to be very small indeed for the effects of dispersal limitation to be apparent in the sample. Conversely, for large microbial communities it will be impossible to distinguish between high immigration rates and the immigration probability being one. This does not mean that immigration will have no affect on the taxa abundance distribution of the community. It just means that the effects are difficult to see in small samples unless they are pronounced.

4.4 Conclusions

The research presented in this chapter demonstrates how mathematical modelling is an indispensable guide to the rational exploration of the microbial world. The huge discrepancy between sample size and the size of microbial communities leave us no option. This is amply demonstrated by the simple, sampling exercise outlined in section 4.1 which clearly demonstrates the dangers of naively extrapolating from small samples. This is important, a proper understanding of the nature of taxon abundance curves is central to the longstanding conundrum of the extent of prokaryote diversity (Curtis & Sloan 2005) and the curves may be (rightly or wrongly) interpreted as reflecting underlying ecological processes (May 1975).

The neutral model deployed in this thesis is simple and can be calibrated. The importance of these two attributes should be emphasised. A model that cannot be calibrated cannot be used to predict and prediction is highly desirable in theoretical microbial ecology. This is because many of the basic patterns in the communities examined remain unknown. Thus, extrapolation from the data and patterns which can be observed is essential in order to make predictions about community structure. These can then be tested using appropriately targeted experimental programmes. The low numbers of parameters deployed in the NCM arises from its conceptual simplicity. It might be argued that the model is too simple to offer any guidance. However, as demonstrated in the previous chapter, the model does appear to be consistent with patterns observed in microbial communities and the theory has been successfully applied to higher organisms (Hubbell 2001b). This does not preclude the possibility of further refinements, or the necessity of rigorous testing. However, it does suggest that it constitutes a sound foundation for the rational exploration of the microbial world.

Although quantifying the immigration events into a microbial system directly is not a viable option there are a number of variables which can be manipulated to give an idea of how the immigration rate, m, affects the systems. For example, a simple suite of laboratory experiments could be conducted with multiple wastewater treatment reactors being fed a common influent, but with different sludge retention times. By examining the community composition and functional stability within each bioreactor the effects of migration may become clearer. A neutral community model would predict that increased migration (and thus shorter sludge retention times) would provide greater stability of diversity and function. That said, there must also be a minimum time for which pollutants must remain in the reactor to avoid them being flushed out before being properly metabolised. The challenge for environmental engineers would be to find this balance and optimum retention time.

The ability to calibrate immigration in samples suggests that a neutral community model at least partly explains community structure. However, to extrapolate to an immigration probability for the community using equation (4.13) requires a knowledge of the size, N_T , of the neutral community. In the sewage works example, is N_T the population of the whole sewage works, in which case immigration would be very low indeed, or are some smaller units, such as flocs, assembling neutrally? Though we do not know the N_T values for the AOB in sewage works we can be confident that there are of the order of 10^6 to 10^8 (Coskuner *et al.* 2005) in a millilitre. It follows, therefore, that the true m values will be very small even in small samples. This may have important implications for the debate on the biogeography of bacteria (Fenchel & Finlay 2005). It is however undoubtedly true that this controversial field would benefit from the rigor that appropriately parameterised mathematical models can bring to a debate.

5 Taxa-Area Relationships for Microbes

One of the most studied concepts in classical ecology is the relationship between an area sampled and the number of distinct species it contains (Arrhenius 1921). Such relationships are used extensively in conservation ecology; it is natural to enquire how diversity is likely to increase or decrease as resources and habitats are either expanded or destroyed.

The most generally cited form of the species area relationship is that of a positive power law, $S \propto A^z$, relating the area in question(A) and the number of species housed within (S). This relationship has been observed for many different groups of macroorganisms, from trees to birds and insects. The values of the z exponent observed for all these disparate lifeforms tend to be of the same order of magnitude, typically between about 0.16 and 0.35. In other words, species are accumulated at a rate proportional to between the third and the sixth root of the area examined.

More recently, this observation has been extended to microbial systems. Published studies on bacteria in salt marshes (Horner-Devine *et al.* 2004a), in the water in bark-lined treeholes (Bell *et al.* 2005a) and many other environments have suggested that this power-law relationship extends to communities of microorganisms as well.

But how could this reported phenomenon be of interest to environmental engineers? Knowledge of how the diversity within a community scales as a function of the size of the system could be invaluable in the rational design of improved wastewater treatment systems. There is growing empirical evidence in support of the intuition of most environmental engineers that there is a link between the composition of groups of microorganisms and the stability or reliability of the functions they fulfil. Even simple indices of community composition such as the overall diversity are being shown to correlate with microbial community productivity and with the spatial heterogeneity of the environment (Kassen *et al.* 2000, figure 5.1). Engineers are beginning to speculate that the differing diversity within wastewater treatment plants of varying design, but treating the same waste (Rowan *et al.* 2003), could well be the cause of observed differences in their functional performances (Curtis & Sloan 2005).

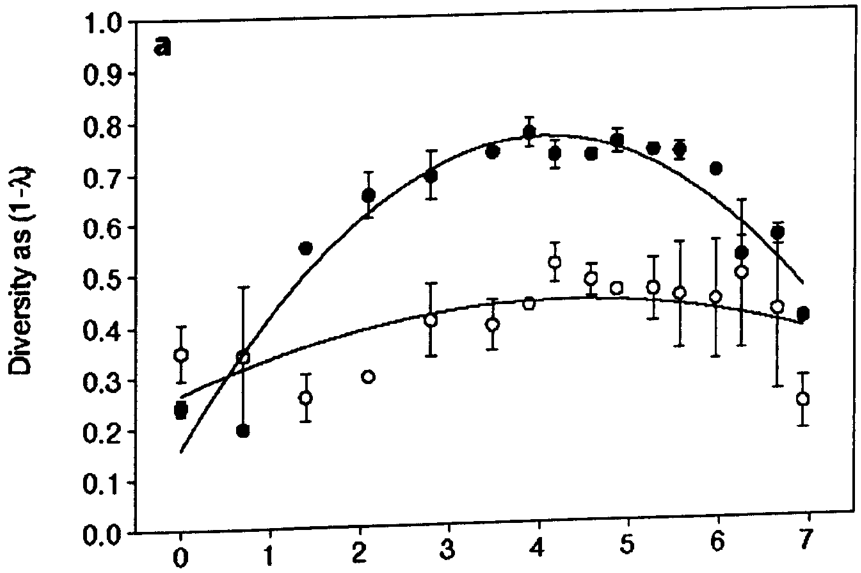


Figure 5.1
Response of *Pseudomonas fluroscens* diversity to nutrient concentration in heterogeneous (solid circles) and homogeneous (open circles) environments. The error bars represent ± 1 S.E. From (Kassen *et al.* 2000).

One of the most common explanations for this correlation between diversity and performance is that it is a reflection of functional redundancy within the system. That is, for the more functionally stable systems, many different taxa are present with similar biochemical functions. Thus, the success or failure of the plant is not dependent upon the ecological success of one single taxon. This belief that function redundancy is vital to engineering functionally stable waste treatment plants is a view is held by some of the most respected theoreticians working in biological engineered systems (Rittmann & McCarty 2001; Curtis *et al.* 2003).

As biodiversity is seen to correlate with the function of such engineered systems, it is critical to have some measure of how the function of plants may be affected if the designs are downsized. The inability to develop smaller scale treatment systems which function reliably has hindered the spread of sanitary water supplies to some developing countries (Saldinger 1992; Newman & Mouritz 1996). Such nations typically lack the infrastructure to support the large plants which are seen to function more consistently. A reliable measure of the relationship between diversity and plant size could well be key to the design of reliable downsized treatment systems.

The recently published studies for microorganisms have, however, so far found vastly differing values for the exponent z in the power-law relationship. These have varied by as much as a whole order of magnitude. Some studies have suggested that the rate of accumulation of microbial taxa with area is similar to that for larger organisms ($z = 0.26$) while others have suggested the rate is much slower, proportional to the fiftieth root of the area sampled ($z = 0.019$). The work presented here, previously published in Woodcock et al. (2006) offers the first quantitative explanation of this huge variance in the observed exponents.

The main achievements outlined in this chapter include:

- Generating theoretical taxa-area relationships for bacterial communities based on some of the commonly cited models for microbial community structure.
- Modelling the effects of sampling upon the exponents observed for the traditional power-law relationship. This may well explain the disparity in the z-value between different published studies.

- Assuming the detection limitations associated with current microbial tools, establishment of conditions for taxa-area relationships to be observable in samples.

5.1 Sampling And Detection Problems Associated with Taxa-Area Relationships

Through the recent advances in the application of molecular methods to microbial ecology, it has become possible to characterise (Torsvik *et al.* 1990) and search for patterns (Green *et al.* 2004; Horner-Devine *et al.* 2004b; Bell *et al.* 2005b) in microbial diversity. However, as previously discussed (section 2.2, chapter 3) most molecular methods analyse small samples from very large, densely populated communities. This disparity between sample size and community size far exceeds that for surveys of plants and animals. For example, assuming no biases, clone libraries of PCR-amplified 16S rRNA genes typically represent a random sample of tens to hundreds of micoorganisms from an environment which may contain as many as 10^9 individuals per gram (Whitman *et al.* 1998). If DGGE analysis is employed, sample sizes are arguably larger. However, this method exhibits a method-dependent threshold in absolute abundance below which organisms will not be detected, which precludes rare species being observed and effectively truncates the sample distribution. For example, Cocolin *et al.*(Cocolin *et al.* 2000) reported that the sensitivity of DGGE was 10^3 cells (typically about 1% of the cells analysed in a sample containing 10^5 cells). Thus, either as a result of very small samples or high detection limits, many taxa remain unseen in the environment (Dykhuizen 1998). Indeed, it has been argued these sampling limitations are sufficiently restrictive that the majority of species are not detected (Curtis *et al.* 2002) and Dunbar *et al.*(Dunbar *et al.* 2002) graphically demonstrate the considerable effort in cloning and sequencing of rRNA genes required to experimentally determine bacterial diversity in a small soil sample. When very high throughput sequencing becomes routinely available to microbial ecologists, complete census of a sample may become possible. However, it is inconceivable that microbial ecologists will ever be able to completely verify patterns at the landscape-scale by a complete census in the way that has been done for tree communities (Condit *et al.* 2002).

In this section, several theoretical taxa-area relationships are considered. For each, the effects of undersampling and of the detection issues encountered with modern laboratory techniques are then modelled. For this study, synthetic microbial communities from homogeneous environments are assembled and samples taken from them. The analysis begins with open, well-mixed, island-like communities of differing size each with similar homogeneous environments; in accordance with the Theory of Island Biogeography larger communities support higher diversity (MacArthur & Wilson 1967; May 1975). In such "well-mixed" communities nothing can be assumed about the spatial correlation in the abundance of taxa. In reality this might arise through the community being truly well mixed, for example, in continually stirred bioreactors (Leclerc *et al.* 2004) or in mixed natural surface waters, or from the spatial structure being obliterated by the sampling procedure (Bell *et al.* 2005b). The communities ranged in size from 10^{10} individuals, which one might expect in a few litres of lake water or tens of millilitres of activated sludge, to 10^{18} individuals; a large lake or wastewater treatment plant. Three assumptions allow plausible synthetic microbial communities to be generated. Firstly, the distribution of taxa abundances in each community is lognormal. Secondly, there is a degree of similarity in the relative abundance distributions for taxa in different sized communities (>10^{10} microorganisms) in the same type of environment; this assumption underlies all theoretical explanations of taxa-area relationships (May 1975; Leitner & Rosenzweig 1997). Thirdly, the lognormal distributions for a particular environment can be crudely characterised using the ratio of two measurable variables, as described in section 2.3, (Curtis *et al.* 2002): the total number of individuals in the community, N_T, which can be confidently measured in samples as the total microbial count, and the abundance of the most abundant members of the community, N_{max}, which can be approximated from clone libraries or more reliably estimated using quantitative molecular methods such as fluorescent *in-situ* hybridisation. This characteristic ratio derived from samples should reflect the community ratio because it is based on the abundance of the most abundant organism (Curtis *et al.* 2002).

The precise nature of species abundance distributions for micro-organisms and each of these assumptions is open to debate (Hughes *et al.* 2001; Ward 2002; Nee 2003). However, provided one accepts that there are likely to be rare species in the microbial community, that the species abundance distribution is uneven and that particular environments, or groups of organisms, will have a characteristic distribution, then the qualitative conclusion of the analysis which follows will hold, irrespective of the precise distribution or parameters (He and Legendre, 2002).

Three different environments have been considered each with its own characteristic N_T/N_{max} ratio: $N_T/N_{max} = 5$, which has been observed in, for example, 16S rRNA gene clone libraries from marine environments (Mullins *et al.* 1995) ; $N_T/N_{max} = 10$, the minimum one might expect in soil (McCaig *et al.* 1999); and $N_T/N_{max} = 25$ which would be more typical of soil (McCaig *et al.* 1999) and has also been observed in anaerobic digesters (Godon *et al.* 1997). Figure 5.1 shows the relationship between community diversity, S, and community size, N_T, within the range 10^{10}-10^{18} individuals for the most diverse environment ($N_T/N_{max} = 25$). Assuming a constant density of organisms in the environment we see that the power law species area relationship, $S \propto A^z$, holds as it does in the less diverse environments (Table 5.1). However, the values of the exponent z (Table 5.1) vary from 0.19 to 0.27; much larger than two of the previously reported values and more typical of large organisms (Rosenzweig 1975). Clearly the precise value of z is reliant on the assumptions about the characteristic taxa-abundance distributions in each environment, but the emphasis here is on how the value of z changes when it is based on small-samples analysed using methods typically used to characterise microbial communities.

Consider the diversity expected in random samples of size 10^6 individuals, which one might reasonably expect in a millilitre of seawater or a milligram of soil (Whitman *et al.* 1998). It is impractical to explicitly generate synthetic communities of 10^{10}-10^{18} discrete individuals for which taxa abundances are lognormally distributed and then repeatedly take very large samples.

As outlined in the previous chapter, under the assumption that $N_S \ll N_T$, it is the case that the joint probability density of taxa abundances with in a sample ($Y = (y_1,\ldots,y_n)$) is Dirichlet distributed. For known local community abundances $X = (x_1,\ldots,x_n)$, $Y|X \sim Dir(N_S x_1,\ldots,N_S x_n)$.

That is,

$$f(Y \mid X) = \Gamma(N_S)\prod \frac{y_i^{N_S x_i}}{\Gamma(N_S x_i)} \tag{5.1}$$

Using MATLAB, theoretical lognormal communities were generated of sizes $10^{10}, 10^{11},\ldots,10^{18}$ individuals for $N_T / N_{\max}$ ratios of 5, 10 and 25. From each community, the above procedure was employed to simulate samples of size $N_S = 200$ and $N_S = 10^6$ individuals. It was then noted how many taxa were present in each sample (*i.e.* how many y_i satisfied $N_S y_i \geq 1$). To simulate a threshold in detection, how many taxa were present at an abundance of at least 10^3 individuals was recorded (counting only $N_S y_i \geq 10^3$). These values were used for calculating the sample z value and the observed z value respectively. The procedure was run for 100 repetitions and least-squares linear regression was used to find z where $\ln S = z \ln A + C$. The appropriate Fisher statistic was then calculated for the regression and the goodness of fit tested at the P = 0.05 level. Additionally, 95% confidence intervals for the exponents were noted. These results are noted in table 5.1.

	$N_T/N_{max} = 5$	$N_T/N_{max} = 10$	$N_T/N_{max} = 25$
Whole community	0.1985 ± 0.0001	0.2415 ± 0.0002	0.2766 ± 0.0003
Complete census of sample: 10^6 microbes	0.0575 ± 0.0039	0.0740 ± 0.0025	0.0747 ± 0.0017
Clone Library; random sample of 200 microbes	-0.0015 ± 0.0070 *	-0.0012 ± 0.0041 *	-0.0049 ± 0.0053 *
Rapid community fingerprinting: 10^6 microbes with a 1% detection limit on relative abundance	-0.0035 ± 0.0073 *	-0.0063 ± 0.0043 *	-0.0050 ± 0.0020 *

- Denotes that the relationship is not statistically significant, therefore, there is no evidence to assume a z value other than 0.

Table 5.1

The exponent and its 95% confidence limits for the power law taxa abundance distribution derived from samples in different environments.

This analysis revealed that a significant power-law relationship holds for analyses based on a complete community census and for samples of size 10^6 (Figure 5.2). However, in each of the environments the z values are significantly reduced when the data are obtained from a small sample (Table 5.1). However, even in a sample of 10^6 individuals it is currently impossible to determine the identity of every individual in a sample. If the analysis is based on samples typically obtained with conventional culture-independent analyses the effective sample sizes relative to the size of the community become very small indeed. Clone libraries prepared from PCR-amplified 16S rRNA genes or functional genes represent a very small random sample from the environment; typically a few hundred organisms. The mean diversity in samples of 200 individuals drawn at random from the environmental samples shows no significant relationship between number of individuals in the community and diversity in the sample (Figure 5.2). The mean z values indicate that they do not differ significantly from 0 and significant taxa-area relationship cannot be discerned (Table 5.1). If more of the genes in the sample are analysed using a community finger-printing (Leclerc *et al.* 2004) method then there will be a threshold below which organisms cannot be detected. Here it has been assumed that all the nucleic acid from a sample of 10^6 individuals is analysed but that sequences with an absolute abundance less than 10^3 are not detected. Again there was no significant relationship between community size and sample diversity for any of the environments (Figure 5.2, Table 5.1).

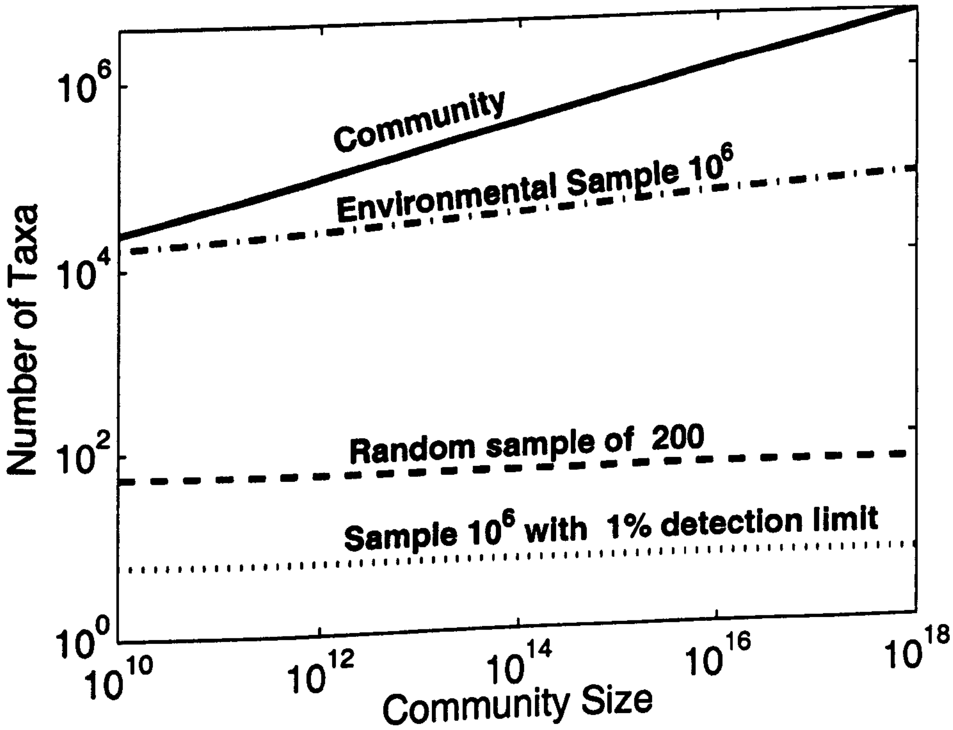


Figure 5.2
Taxa-area relationships for a homogeneous environment with a lognormal taxa-abundance distribution defined by $N_T / N_{\max} = 25$. The correlation is statistically significant at the P=0.05 level for the whole community and in a complete census of environmental samples. The very low gradient in the small samples is not significant.

Until now only random samples from well-mixed Island-like communities with homogeneous environments have been considered. In these cases each individual in the community is equally likely to form part of the sample. This makes it conceptually and practically difficult to investigate how diversity changes within successively smaller parts of the community (nested taxa-area relationship)(Leitner & Rosenzweig 1997) on the basis of small samples. However, in many environments, such as soils, the spatial distribution of distinct microbe communities will remain relatively fixed in time which opens up the possibility of determining the nested taxa-area relationship based on the decay in similarity with distance between samples (Harte *et al.* 1999), rather than solely the diversity in the sample. However, for this to be possible the presence or absence of taxa must be spatially correlated, either as a result of environmental gradients or by, for example, dispersal limitation (Hubbell 2001). Nested taxa-area relationships will exist in homogeneous spatially uncorrelated environments, but are, perhaps unsurprisingly, impossible to determine based on similarity decay methods. To reinforce this observation the nested-taxa area relationship in a spatially fixed but uncorrelated community of 10^{16} individuals is examined; the number of microbes one might expect in, for example, 10 tonnes of soil. Again, the environment is assumed to have a characteristic lognormal distribution of taxa-abundances defined by $N_T/N_{max} = 25$.

The 10^{16} individuals are distributed uniformly in space on a disc with unit radius and the whole-community nested taxa-area relationship is determined from a complete census of the organisms lying within concentric discs of increasing area; the smallest disc, in the centre, comprising only 10^6 individuals. Subsequent samples of equal size are taken from the annuli that lie between successive concentric circles. The Sorensen similarity index (Harte *et al.* 1999) between the sample from the central disc and those in each annulus is then determined. The distance of the samples from the centre of the unit disc is assumed to be approximately the area weighted average radius of the annulus; $\sqrt{(r_1^2 + r_2^2)/2}$, where r_1 and r_2 are the radii of inner and outer boundaries of the annulus from which the sample was taken. Figure 5.4 shows the similarity of samples as a function of distance for a typical realisation of the community. As expected, no decay in similarity with distance is apparent even if a complete census of the samples were possible.

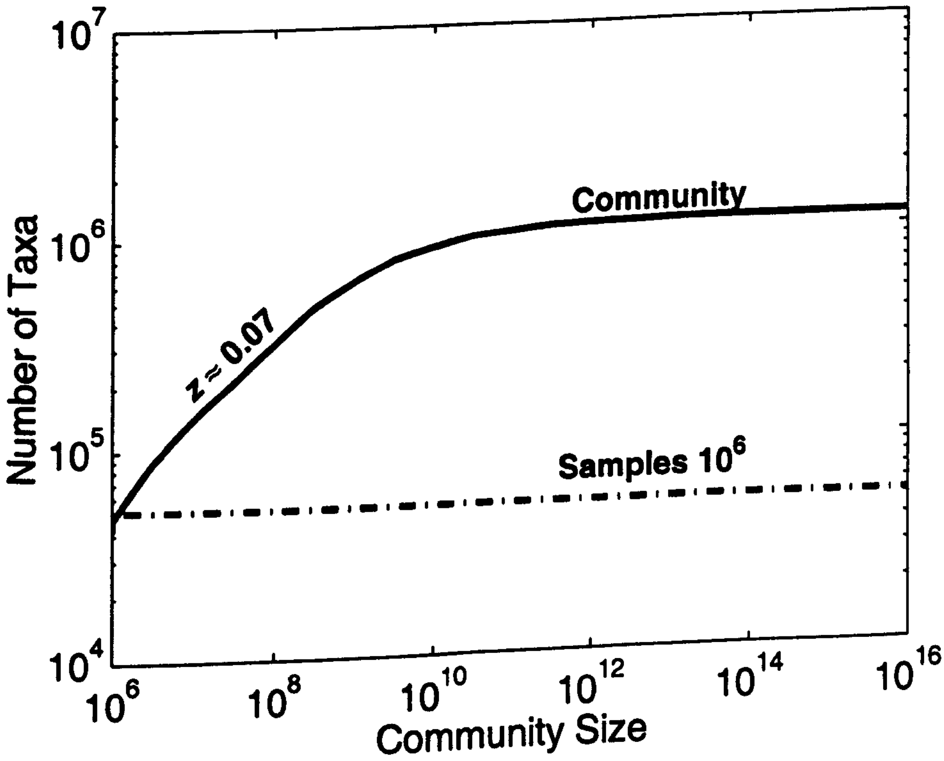


Figure 5.3
Nested species area relationship for the whole community in a homogeneous spatially uncorrelated environment.

Initially, the exponent of the nested taxa-area relationship is approximately 0.07 which is similar to that for large organisms (Rosenzweig 1975), it levels off at around 10^{12} organisms when most taxa have been encountered. There is no reason to expect similarity in samples to decay with distance in a homogeneous spatially uncorrelated environment, and this is borne out in samples the samples of size 10^6 by simulation (Figure 5.3). Using a typical distance-decay method for ascertaining species-area curves (Harte *et al.* 1999) this translates into a completely flat taxa-area relationship (Figure 5.4).

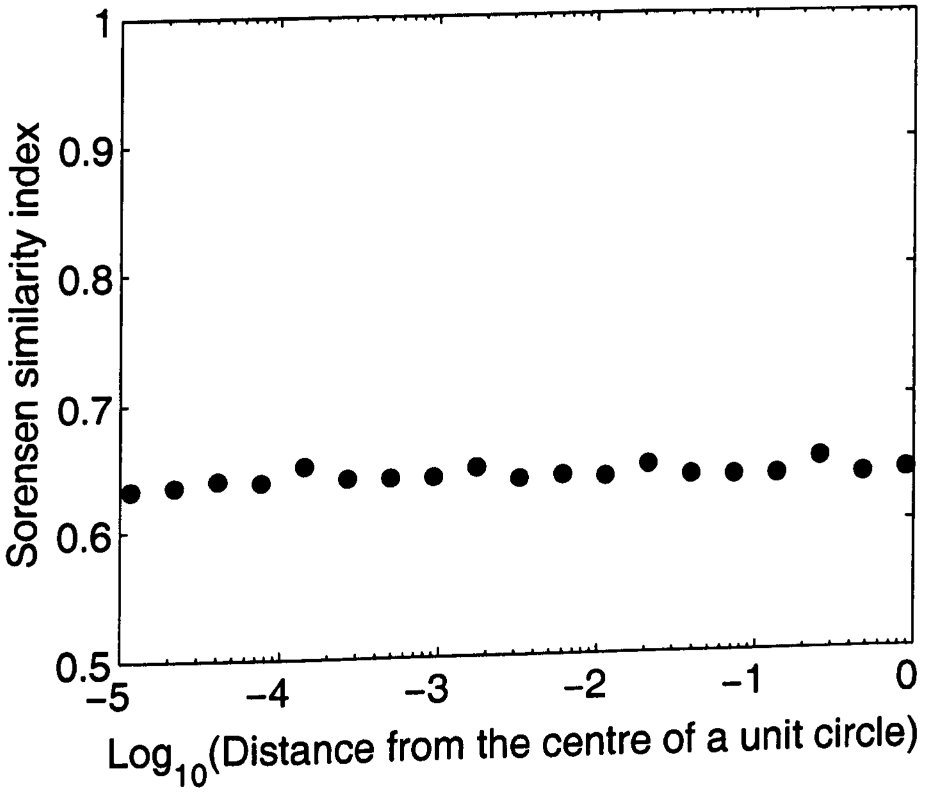


Figure 5.4
Sorensen similarity indices for samples of varying distance from the centre of a unit disc. Note that similarity does not change with distance between samples, therefore the community taxa-area relationship cannot be inferred.

5.2 Conclusions

The observation of taxa-area relationships (Green *et al.* 2004; Horner-Devine *et al.* 2004a; Bell *et al.* 2005a) is an important break-through in microbial ecology. These observations have been possible because of the advent of molecular methods. However, the relatively narrow dynamic range of these methods means that it is difficult to simultaneously detect common and rare members in a microbial community. Thus, when conducting broad-scale analyses of microbial communities the accumulation of rarer taxa that dictates the value of z in taxa-area relationships is difficult to detect. If taxa-area relationships are ubiquitous, their true nature will almost always be disguised, or even hidden, as a result of inherent limitations of the measurement methods.

The sampling issues highlighted by this analysis are of profound importance for those seeking patterns in microbial ecology and potential applications in environmental engineering. Taxa-area relationships may be difficult to observe, even if they exist. What has been demonstrated here is that, under the additional assumptions of there being no shift in community structure with area and the top ranked abundance staying constant across sites, taxa-area relationships are almost impossible to detect. For such studies as those of Green *et al.* (2004) and Bell *et al.* (2005a), who noted high z-exponents, some further mechanism must be at work, one which produces a correlation between the abundances of the top few ranked species and the area studied. Without this additional factor, they would likely have been unable to observe the phenomena they did.

6 Evidence of Neutral Community Assembly

The previous chapters of this thesis have been concerned with the development of NCMs and their possible applications in engineered biological systems. So far, most of the work presented has been largely theoretical and the predictive power of such models has not been fully considered. In this chapter, the research is expanded upon and compelling evidence presented of neural community assembly in naturally occurring microbial communities. Furthermore, it is demonstrated that NCMs may offer answers to various problems facing environmental engineers.

One key question for the engineering community concerns the link between system sizes and diversity. Knowing at which scales plants should be functionally stable could be important in the design of wastewater treatment systems. Thus, a good understanding of the taxa-area relationships for microbes is of the utmost importance.

However, as the previous chapter demonstrates, the biases and sampling effects inherent to current laboratory tools may well distort or even obscure these scaling effects completely. Here, NCMs are invoked to offer a plausible explanation as to why in some cases such taxa-area relationships should be readily observed in experimental studies and why in other cases, the phenomenon may remain undetected.

Additionally, a comprehensive test of how well a NCM can describe taxa abundance distributions across a wide range of scales. By re-analysing a recently collected dataset (Bell *et al.* 2005) for water borne bacteria living in treeholes in Beech trees in the same woodland, the predictive power of NCMs is tested across physically and chemically similar communities whose sizes span three orders of magnitude. The result is the strongest test yet of the Neutral Theory, either in classical or microbial ecology and forms the basis of a paper to be published in the near future (Woodcock et al. 2007, In Press).

The main achievements outlined in this chapter include:

- Demonstration that Neutral Community Models offer a solution to the problem of detecting taxa-area relationships in small samples. Specifically, it is noted that there are certain conditions and parameter values for which detection of this scaling phenomenon is a simpler task than for others.
- Calibration and validation of the NCM for the dataset first published in Bell *et al.* (2005). By fitting one set of parameters for the smallest site, predictions were generated for the remaining 28 communities. Of these, 26 were found to be statistically significant fits at the 5% level.

The work presented forms the basis of a paper published in Ecology Letters (Woodcock *et al.* 2006) as well as a forthcoming paper in FEMS Microbiology (Woodcock *et al.* 2007 In Press).

6.1 Taxa-Area Relationships Predicted by Neutral Community Models

In the previous chapter, it was demonstrated that the sampling and laboratory techniques employed in microbial community analysis have the potential to severely distort the picture of the z-exponent in the power law taxa-area relationship. Furthermore, for a number of lognormally distributed communities each with the same $N_T / N_{\max}$ ratio, it was shown that the exponent could appear to be more than an order of magnitude smaller in samples than in the real system. This was the case for both nested and island-like taxa-area relationships.

How then, can recently published studies of microbial communities observe exponents similar in magnitude to those found in classical ecology? For example, Bell et al. found a z-value of 0.26 for bacteria inhabiting small water-filled treeholes in a UK forest. This is a remarkable dataset because it offers a perfect analogue to the islands used in the Theory of Island Biogeography (MacArthur & Wilson 1967).

The treehole dataset used throughout this chapter provides the perfect testing ground for NCMs; insular communities of different sizes housed in similar ecosystsm. Samples were taken from 29 rainwater filled, bark-lined holes, each of which housed a small ecosystem. The range of volumes of these habitats spanned three orders of magnitude; the smallest was a mere 50ml, the largest 18,000ml. Bell et al. (2005) reported that bacterial species richness increased with treehole volume in a manner that could be modelled using a single power law relationship which hints at some consistent process of community assembly. Physically and chemically, the bacterial communities shared a great deal in common; they were all supported by similar nutrients (decaying leaf litter), relatively stagnant, but subject to invasion events from either airborne or rainwater borne microorganisms. The greatest geographic distance between any two trees in the study was around two miles.

The analysis contained in chapter 5 suggests that there must be a significant change in the community structure of the more abundant organisms between tree-holes. But the holes are all in the same species of tree in the same forest and the water in each tree hole was stirred before sampling, which lessens the likelihood of the community structure differing as a result of environmental factors that correlate with volume. Thus some other factor must be at play.

One possible explanation for this may be offered by applying the NCMs presented earlier in this thesis. By considering the differing effects of immigration on communities of different sizes, such models can offers a self consistent argument as to why a taxa-area relationship may well be derived for these small insular communities. Although NCMs perhaps do not provide a literal description of community assembly, they do offer the opportunity to investigate the role of immigration in isolation from all other factors.

As discussed in section 2.3, as the $N_T m$ parameter which defines the shape of the species abundance distribution for a NCM decreases (figure 6.1), the distribution becomes more negatively skewed. This implies that, independent of any environmental factors, in a set of different sized islands subject to the same immigration rate the expected abundance of the relative abundance of the top few ranked (by abundance) taxa increases as island size decreases. Such models suggest that where immigration has an important role in determining community structure, small samples analysed using culture-independent methods can yield a much closer estimate of the true taxa-area relationship, despite the lack of knowledge on how rare taxa are accrued. In this section of the thesis the NCM is crudely applied to the treehole dataset (Bell et al. 2005) to demonstrate the generic mechanisms by which the taxa-area relationship can be revealed in dispersal limited systems. In the next section a more rigorous application of the NCM to the same dataset is used to calibrate and validate the model.

To apply a NCM model estimates of the number of individuals in each community are required. Additionally, an estimate the probability, m, that new individuals in the community result from immigration rather than from reproduction in the local community is needed, along with an estimate of the fundamental biodiversity number, θ, which determines the diversity of the metacommunity that supplies immigrants to the tree-holes.

For microbial communities, these cannot always be easily estimated. However, for the communities in the Bell et al. study, assuming that the density of organisms in the tree holes is of the order 10^5 per ml (van der Gast 2005), the population sizes ranged from approximately $10^{6.5}$ to $10^{9.5}$. The ambient density of organisms in the air outside the tree holes was of the order 10^3 per m^3 (Harrison et al., 2005) as compared to 10^{11} per m^3 in the tree hole fluid. This is an enormous difference in density which makes it is difficult for the per capita atmospheric deposition to be very high in the absence of some very turbulent mixing between the air within and outside the tree hole. Assuming relatively stagnant air in the tree hole the flux of organisms across the water surface boundary is likely to be very low indeed in comparison with population turnover due to local reproduction and deaths through, for example, predation. Thus the different densities of organisms in different environments might mean that whilst microorganisms can disperse freely (Fenchel and Finlay, 2005) the per capita immigration rates into some environments are low. Immigration will be punctuated by events like stemflow and animal foraging, however, averaged through time the immigration probability will be low. Thus, a somewhat arbitrarily selected value of 10^{-6} is used here. A biodiversity number of 8 yields an N_T/N_{max} value, for a very large community, of 4 which has been observed in aquatic environments (Mullins et al., 1995).

For these parameters in the model, figure 6.1 shows the ranked abundance distributions for 10 communities within the tree-hole size range. Figure 6.2 shows the ranked abundance distribution of taxa in 5ml samples and figure 6.3 shows the taxa-area relationship for the entire community, for the samples and for molecular community finger printing with a detection threshold of 10^3 individuals, typical for DGGE analysis. It is apparent in figure 6.2 that the increasing evenness of the species abundance distribution that occurs with increasing volume is sufficient to be manifest as an increasing number of bands detected on a DGGE gel.

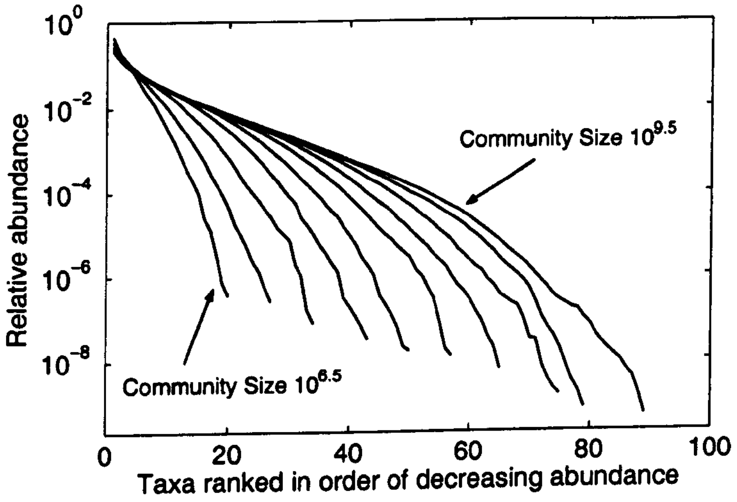


Figure 6.1

Ranked species abundance distribution from a Neutral Community Model for a range of community sizes, with immigration probability 10^{-6} and biodiversity number 8.

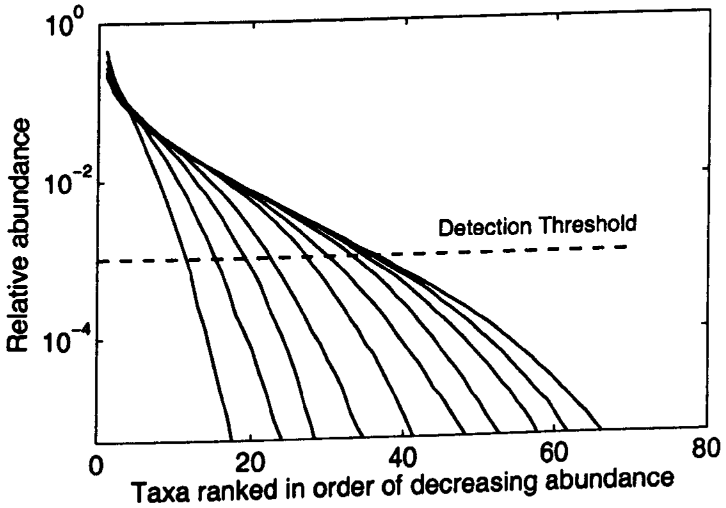


Figure 6.2

Ranked species abundance distribution from a Neutral Community for 5ml samples, with immigration probability 10^{-6} and biodiversity number 8.

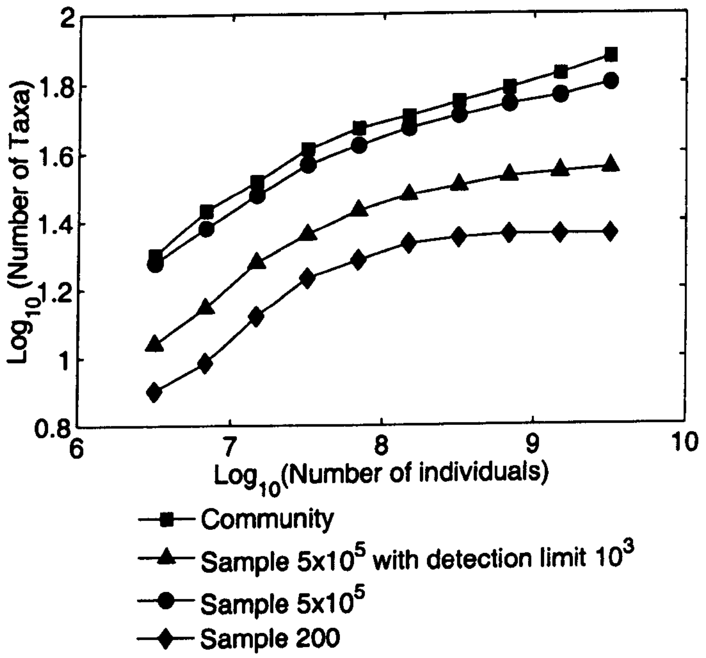


Figure 6.3

Taxa abundance distribution identified using random samples of differing sizes.

This analysis does depend upon the estimates used for the parameters. In particular, it should be noted that there is a lack of empirical evidence to support the immigration rate of 10^{-6}. However, as a rough rule of thumb, the evenness of the taxa abundance distribution in the local community is seen to be sensitive to community size while $N_T m$ < 10000. This difficulty in observing scaling effects for larger systems is consistent with the scaling function displayed in figure 4.2. For systems with a higher $N_T m$ parameter, a far smaller shift occurs in the top few ranked abundances. Also, as discussed in chapter 4, the sample sizes themselves can make what small changes occur extremely difficult or even impossible to detect. For such communities, the effective immigration rate $\tilde{m}$ would always appear close to unity.

This phenomenon of increasing evenness of the taxa abundance distribution with increasing volume could well be key to Bell *et al.* (2005) observing a taxa abundance distribution in small samples. There are alternative explanations such as that of Green and Ostling (2003), who demonstrate that this could occur if taxa-abundances are spatially correlated. However, by invoking the simple neutral community model, a plausible ecological mechanism for the phenomenon occurring in what are ostensibly homogeneous islands can be found. Having demonstrated the generic mechanism which allows microbial taxa-area relationships to be revealed in dispersal limited systems, it remains to validate whether this is indeed occurring.

6.2 Calibration and Validation of Neutral Community Models

Much of the interest in NCMs stems from the fact that such beguilingly simple models can theoretically reproduce some of the fundamental patterns in nature which ecologists have been trying to explain for decades. However, NCMs have only ever been fitted using taxa-abundance distributions from single sites (Volkov *et al.* 2003) or at one scale (Sloan *et al.* 2006), (section 3.3) and parameter values have been calibrated on a case-by-case basis. Since NCMs predict a malleable two parameter taxa-abundance distribution, this is a weak test of neutral community assembly and, hence, of the predictive power of NCMs.

In this section, a NCM is applied with a single set of parameters to predict the taxa-abundance distributions and taxa-volume relationship observed in the waterborne bacterial communities studied in Bell *et al.* (2005), whose sizes spanned three orders of magnitudes. This validates the simple quantitative ecological mechanism of dispersal limitation proposed in the previous section and also demonstrates the predictive power of NCMs.

In most previously published applications of neutral theory, the model parameters have been selected to minimise the difference between observed and predicted taxa-abundance distributions. The merit of NCMs over and above other hypotheses on the formation of biological communities is then argued on the basis of (often small differences in) a goodness of fit statistic for calibrated taxa-abundance distributions (Chave *et al.* 2002; McGill 2003; Volkov *et al.* 2003). These arguments can seem rather arcane when there has been little attempt to validate the models (Harte 2003). In addition, microbial ecologists are precluded from the debate because, for most environments, only a small fraction of the diversity can be experimentally defined, as discussed in the previous chapters. Despite the advances in molecular methods for characterising naturally occurring microbial communities *in situ* the disparity in scale between sample and community size and some inherent limitations of the methods conspire to make a purely empirical definition of a taxa-abundance distribution at a single site very difficult. The frequency-p_i method presented in chapters 3 and 4 is one attempt at circumventing this problem for sample data from similar microbial communities. Nonetheless, calibrating an NCM at one site or one scale is not a convincing endorsement of the model's underlying assumptions and many alternative models could potentially reproduce either the taxa-abundance distributions(McGill 2003) or the abundance-frequency relationships observed. Part of the fascination with NCMs is their potential to predict the biogeography of groups of organisms as a function of key variables: the immigration rate, community size or the distance between samples. Finding examples of where such predictions hold true or fail will yield greater insight on the utility of such a parsimonious description of community assembly as an NCM than curve fitting.

The distributions of the relative abundance of taxa in the samples was not reported in Bell et al. but are used here (Figure 6.4). What is immediately striking from these data is just how dramatically the shape of the taxa abundance distributions change between treeholes, with large communities exhibiting a much more even distribution than small ones, which is in broad agreement with the effects of dispersal limitation discussed in the previous section. Given the proximity of tree holes and the similarity of their environments, these data provide ample opportunity to test the reasonable null hypothesis that the tree holes house distinct homogenous island-like communities that are neutrally assembled from a single metacommunity with a consistent rate of random immigrations into each tree hole.

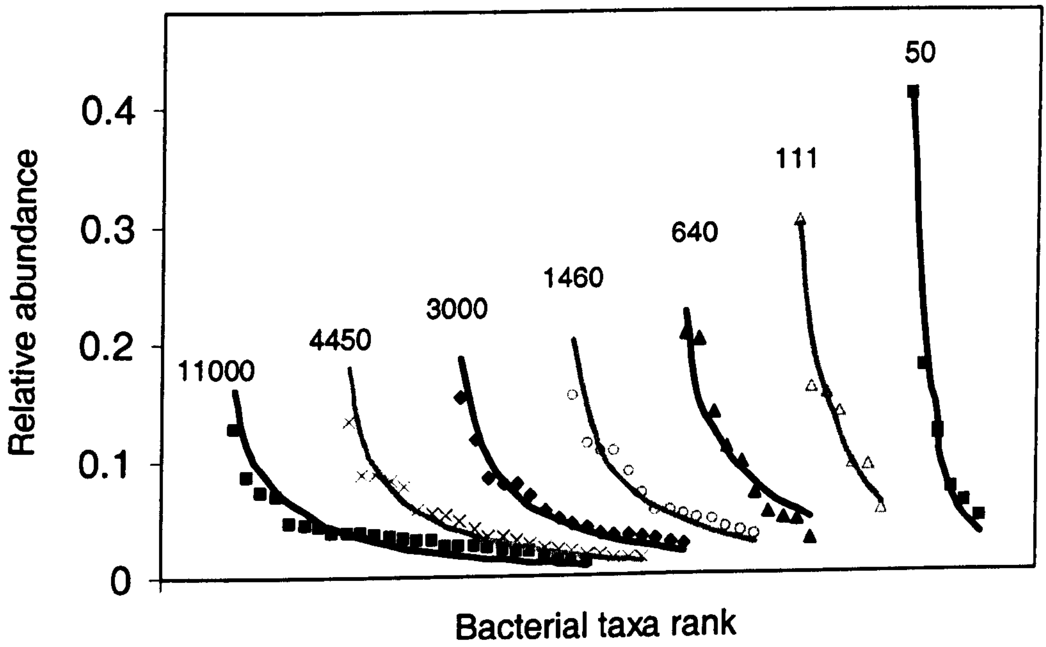


Figure 6.4

Ranked taxa abundance distributions for a selection of 7 of the 29 treeholes ranging in volume from 11000ml to 50ml. Lines represent taxa abundance distributions predicted by neutral model with $m=10^{-6}$ and $\theta=15$ calibrated using data from the 50ml treehole.

Because of detection limitations inherent to the DGGE analysis, in the initial study, only the top few ranked taxa were observed at each site, so the abundances in the dataset were normalised relative only to the total abundances of these most common taxa. Accordingly, in the analysis all the predicted abundances were normalised relative to the detected number of taxa in the dataset.

Initially, the model was calibrated using only the observed taxa-abundance distribution for the smallest (50ml) treehole. By selecting values for the parameters θ and m, realisations of the metacommunity were generated.

Then, as outlined in sections 3.2 and 4.2, theoretical local populations and sample populations were produced via appropriately normalised Gamma variables. After running simulations with many different parameter pairs, it was found that the least-squares best fit to be obtained with $\theta = 15$ and $m = 1.0 \times 10^{-6}$. These parameters were then used to predict the taxa-abundance distributions for all the remaining treeholes. Comparisons between observed and simulated taxa abundance distributions for a selection of the tree holes are shown in Figure 6.4.

Applying these parameters, the resulting prediction for each site was tested individually to see if the neutral model fitted the data at the 5% significance level. As no simple analytic method is available to find the expected abundances of each ranked taxon in a NCM, simulations were employed to find these. 1000 repetitions of the neutral model were run for sites of the volumes of each of the 29 treeholes using the selected parameter pair and the mean values of the ranked abundances were assumed to closely approximate their expectations.

Using these expected abundances, 500 independent repetitions were calculated and Pearson's Statistic for goodness of fit was calculated for each

$$\sum_{i=1}^{n} \frac{\left(E_{(i)} - x_{(i)}\right)^2}{E_{(i)}}$$

where $E_{(i)}$ is the expected abundance of the i^{th} ranked taxon and $x_{(i)}$ is its abundance in the simulation or observed dataset. A p-value was then estimated from the proportion of these 500 trials which produced a goodness of fit statistic greater than that calculated for the observed data.

For those sites for which a p-value less than 0.05 was obtained, the null hypothesis of a neutral model with the parameter pair $(15, 10^{-6})$ was rejected.

Treehole Number	Volume (ml)	p-value
1	360	0.076
2	3250	0.324
3	1700	0.500
4	750	0.280
5	18000	0.020*
6	180	0.620
7	640	0.454
8	4450	0.612
9	3600	0.152
10	3150	0.146
11	2250	0.326
12	1800	0.650
13	1250	0.154
14	60	0.480
15	1950	0.122
16	2850	0.214
17	2225	0.956
18	900	0.228
19	11000	0.158
20	1460	0.588
21	50	0.956
22	3000	0.674
23	140	0.342
24	220	0.094
25	111	0.808
26	350	0.466
27	1200	0.068
28	3000	0.736
29	600	0.040*

* Not statistically significant

Table 6.1

Estimated p-values for the goodness of fit using a NCM calibrated against the smallest site, treehole 21. The parameter pair used was $(\theta, m) = (15, 10^{-6})$.

Accordingly, hypothesis testing at the 5% significance level suggested that for 27 of the 29 treehole communities there was no evidence to reject the neutral model. The predicted species richness in each treehole closely matched those observed and, consequently, the neutral model reproduces the taxa-volume relationship reported in Bell et al (2005) (Bell *et al.* 2005) for samples from the community (Figure 6.5). When the parameter pair was calibrated using all the data from all the treeholes, it was found that the best fit (sum of root mean square errors for each tree hole) for the entire dataset was given by $\theta=20$, $m=1.5\text{x}10^{-6}$. Testing at the 5% significance level again gave no reason to reject the neutral model with these parameters in the same 27 treeholes. The goodness of fit was insensitive to changes in the parameters within the relatively small range: $15 < \theta < 25$ and $5\text{x}10^{-7} < m < 5\text{x}10^{-6}$.

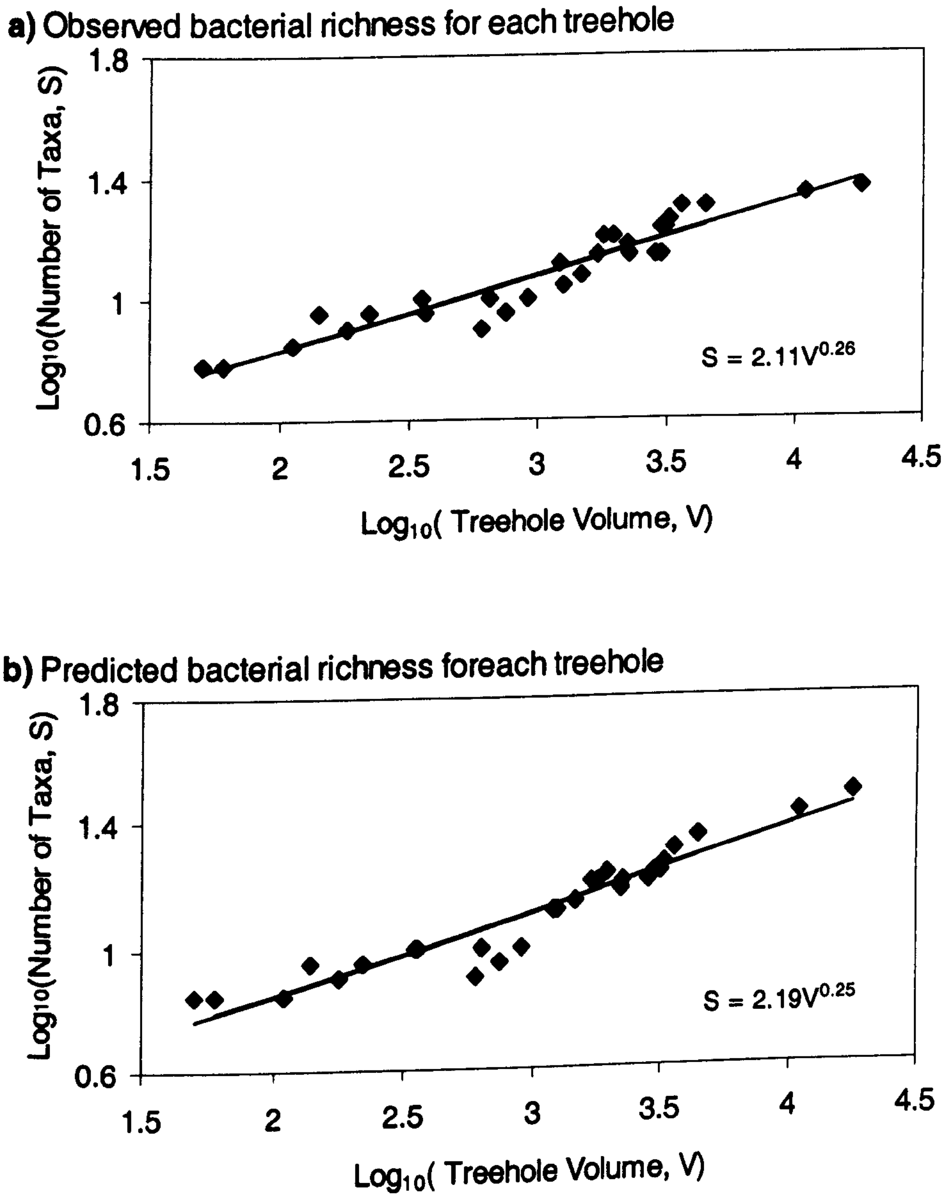


Figure 6.5

a) The observed bacterial richness in all 29 tree holes. The solid line represents the power-law relationship, $S = 2.11V^{0.26}$, fitted using linear regression.

b) The bacterial richness predicted by the neutral model with $\theta = 15$ and $m = 10^{-6}$ calibrated using the taxa-abundance distribution of the smallest treehole. The solid line represents the power-law relationship, $S = 2.19V^{0.25}$, again fitted using linear regression.

6.3 Conclusions

In the previous chapter, it was demonstrated that, under the assumption that the abundances of the top ranked taxa are not correlated with the area sampled, the true nature of taxa-area relationships is virtually undetectable using current molecular methods for community analysis, just as some studies had found (Horner-Devine *et al.* 2004). There are, however, a number of published reports (Bell *et al.* 2005) which have indeed found a strong taxa-area relationship with exponents of the order of those observed for larger organisms $z \approx 0.25$. There had been no previously published explanations offered as to why these seemingly contradictory observations may have arisen prior to the analyses in this and the previous chapter. By invoking a simple neutral community model, a potential solution can be found to this apparent contradiction between different studies.

Furthermore, it is shown that a neutral community model can be calibrated on a single site and the same parameters used across similar communities of vastly different sizes to predict community composition with great accuracy. The success of the neutral model in explaining the different taxa abundance distributions in tree holes, whose sizes vary over three orders of magnitude, without the need to change any parameters constitutes the strongest evidence, so far, that random reproduction, death and immigration play a significant role in shaping bacterial community structure. It suggests that at least some bacterial communities are dispersal limited and, therefore, challenges the perspectives held by some commentators that global dispersal of microorganisms prevents them having a biogeography (Fenchel 2003) and that microbial population sizes are sufficiently large to preclude local stochastic extinctions (Fenchel & Finlay 2005).

Naturally occurring communities of microorganisms are vital to life on Earth and are of profound practical significance in agriculture, medicine and engineering. Describing patterns in microbial communities is, therefore, important but not as important as explaining why the patterns form. Quantitative theories of microbial community assembly could allow the composition and dynamics of naturally occurring communities of microorganisms to be predicted and manipulated to the benefit of engineering, medical, veterinary and environmental science. There may be many qualitative alternative explanations of the patterns in the bacterial data presented here but, as Harte (2004) suggests, theories are of most interest when the ratio of the number of predictions that they make to the number of assumptions and adjustable parameters is low. It has been shown that a simple two parameter neutral community model, calibrated at one site, can predict patterns in bacterial biodiversity and biogeography over many different sites and scales.

7 Discussion

As the demand for cost-effective treatment strategies for both industrial and domestic waste grows ever greater, the pressure is on environmental engineers to find improved approaches to the problem. Many of the currently employed reactor designs are already a century old (Ardern & Lockett 1914; Metcalf & Eddy 1914), and are founded upon empirical rules which mean that changes in design are slow and often risky and failure, when it occurs, is often inexplicable(Curtis *et al.* 2003).

It is now widely accepted (Rittmann *et al.* 2006) that strategies founded upon a more complete understanding of the composition of the microbial communities employed would lead to the improved function of bioreactors. Consequently, environmental engineering has been rebranded as environmental biotechnology to reflect the increasing collaboration between environmental engineers and microbial ecologists.

The vast wealth of microbial molecular methods developed in the past two decades is offering ever greater and more detailed glimpses (Moffett *et al.* 2000; Daims *et al.* 2005) into the microbial communities relied upon for such environmental biotechnologies. As more detailed datasets become available, the belief is that rules governing microbial community assembly can be formulated and theories similar to many of those in classical ecology can be postulated. Just as the application of the laws of Newtonian physics to structural mechanics instigated a revolution in structural design, the development of rules for microbial community assembly may well have a similar impact on waste treatment strategies.

However in classical ecology, the communities, and the patterns therein, are relatively easy to observe, although the ultimate causes of those patterns are obscure. In this context, complex models can give invaluable insights into the underlying mechanisms that drive community composition and structure (Loreau, 2004; Tilman, 2004) even though such models are very difficult to parameterise. By contrast, microbial communities remain relatively difficult to observe. The very best molecular techniques yield only partial and fleeting glimpses of the communities to which they are applied. In this context complex and un-parameterised models have the potential to mislead, because we have a poor grasp of the reality they seek to describe. Consequently, for the time being at least, a simple model which can be parameterised, like the neutral community model presented in this thesis, may be the better guide to the microbial world. A model does not need to be mechanistically correct in every detail to be useful (Harte, 2003).

The development of theories to describe microbial community assembly is still at an early stage (Curtis et al., 2002; Green et al., 2004; Horner-Devine et al., 2004). The work presented here has perhaps simply extended, formally and quantitatively, the principles outlined in The Theory of Island Biogeography (MacArthur & Wilson 1967) to microbial systems. Nevertheless, it is worth remembering, that the Theory of Island Biogeography transformed the understanding and application of ecology. An analogous transformation in microbial ecology is long overdue. Such a transformation would mean that the composition and dynamics of naturally occurring communities of microorganisms could be predicted and manipulated to the benefit of not only environmental engineering problems, but also a number of those in medical, veterinary and environmental science. The theory described here may, either in its own right or as a foundation for more sophisticated approaches, make a contribution to this goal.

8 Future Research Questions

The ultimate goal for this area of research is the development of a suite of laws and concepts which can be used to describe and predict the assembly and structure of microbial communities. Such laws would then give environmental engineers greater opportunities to modify and improve the design and efficiency of waste treatment systems.

The majority of the research in this thesis takes neutral community models (NCMs) as the starting point for this goal. As previously discussed, they represent an attractive null model, although may well have limitations which need to be overcome. The following sections of this chapter briefly discuss some possible future lines of research which have not yet been fully pursued.

8.1 Examination of Neutral Community Dynamics

As discussed in section 3.1 as well as being a route towards the steady-state distribution, equation 3.9 can be translated into an Ito stochastic differential equation(Kloeden and Platen 1999).

$$dx_i = M_{\delta x_i} dt + \sqrt{V_{\delta x_i}} dw \tag{8.1}$$

where w is a standard Wiener process. Employing the Euler-Maruyama method, this can be reduced to the sum of a number of small discrete steps.

The time interval is first broken into steps of size dt, and then realisations of a normal distribution generated $dw_0, dw_1, \ldots$ such that all dw_js are independent, identically distributed variables where each $dw_j \sim N(0, \Delta t)$. Setting $X_0 = x_i(t = 0)$ and then defining

$$X_{j+1} = X_j + M_{\delta x_i} \Delta t + \sqrt{V_{\delta x_i}} dw_j (\Delta t) \tag{8.2}$$

a vector of points can be generated that forms an approximate solution to the stochastic differential equation.

This equating of the expected speed of the dynamics could well provide a route to understanding at what scale communities can be assumed to be homogeneous. The neutral model described in this thesis assumes that the whole community studied, whether a wastewater plant or an estuary, is homogenous within those boundaries. However, some preliminary work has been done on the dynamics and it has been found that with this assumption, the communities change unfeasibly slowly, around 1000 or more times less rapidly than is commonly observed. This had led some observers (Nee 2005) to be highly critical of neutral theories. There is, however, one other observation. By considering the communities as patches of smaller sub-communities, each itself completely homogenous, the dynamics of the whole community would be altered. It remains a currently unanswered question as to whether studying the dynamics of a real system compared with the theoretical dynamics for systems using different patch sizes would give a better understanding at quite what scale communities can be assumed to be homogeneous.

8.2 Near-Neutral Community Models

One of the main strengths of NCMs is the ease of their analysis. The assumption of homogeny between all individuals in a community circumvents the need from detailed measurements of many kinetic parameters, and so massively simplifies any calculations which may be needed. However, this may be over simplification. Failure to accommodate any differences in the fitness of a species to a given environment could render the model useless if applied to very different environments. For example, in the extreme case of applying an NCM to a warm saline water body and a cold fresh water body the same community structure would be generated if a similar metacommunity is assumed. This might be considered an abuse of the NCM because the metacommunities for the two environments would be radically different. However, even small variations in fitness in the local community may have an effect on the biodiversity that overwhelms the stochastic demography involked in a neutral model. There are many different approaches to building such inter-species variations into the framework presented in this thesis. One of the simplest formed preliminary work on this topic which was published as part of a paper in Environmental Microbiology (Sloan et. al 2006). For that approach, a simple parameter, α_i, was added to the model which could be set to define the relative advantage or disadvantage of each species to compete within a local community.

Further work in this area would be to examine whether calibration of the distribution of α_is and hence calculation of the taxa-abundance distribution for near-neutral communities would be feasible. Furthermore, just as resources have been incorporated into some classical wastewater treatment models it may well be desirable to build in the competitive abilities of each taxon as a function of the concentration of one or more resources.

9 References

Amann R.I., Binder B.J., Olson R.J., Chisholm S.W., Devereux R. & Stahl D.A. (1990a) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology*, 56, 1919-1925.

Amann R.I., Krumholz L. & Stahl D.A. (1990b) Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic and environmental studies in microbiology. *Journal of Bacteriology*, 172, 762-770

Ardern E. & Lockett W.T. (1914) Experimentation on the oxidation of sewage with the aid of filters. *Journal Soc. Chem., Ind.*, 33, 523-539

Arrhenius O. (1921) Species and area. *Ecology*, 9, 95-99

Bell G. (2000) The distribution of abundance in neutral communities. *American Naturalist*, 155, 606-617

Bell T., Ager D., Song J., Newman J.A., Thompson I.P., Lilley A.K. & van der Gast C.J. (2005a) Larger Islands House More Bacterial Taxa. *Science*, 308, 1884

Bell T., Newman J.A., Silverman B.S., Turner S.L. & Lilley A.K. (2005b) The contribution of species richness and composition to bacterial services. *Nature*, 436, 1157-1160

Boggs B.A. & Chinault A.C. (1997) Analysis of DNA replication by fluorescence in situ hybridization. *Methods*, 13, 259-270

Borneman J. & Triplett E.W. (1997) Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation. *Appl. Environ. Microbiol.*, 63, 2647-2653

Boswell M.T. & Patil G.P. (1971) Chance mechanisms generating the logarithmetic series distribution used in the analysis of number of species and individuals. In: *Statistical Ecology* (eds. Patil GP, Pielou EC & Wates WE), pp. 99-130. Penn State University Press, Philadelphia, PA

Browne R.A. (1981) Lakes as islands: biogeographic distribution, turnover rates, and species composition in the lakes of central New York. *Journal of Biogeography*, 8, 75-83

Bunge J., Epstein S.S. & Peterson D.G. (2006) Comment on "Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil". *Science*, 313, 918c

Caswell H. (1976) Community structure: A neutral model analysis. In: *Ecological monographs*, pp. 327-354

Chandler D.P., Fredrickson J.K. & Brockman F.J. (1997) Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Molecular Ecology*, 6, 475-482

Chao A., Chazdon R.L., Colwell R.K. & Shen T.-J. (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, 8, 148-159

Chave J., Muller-Landau H.C. & Levin S.A. (2002) Comparing classical community models: Theoretical consequences for patterns of diversity. *American Naturalist*, 159, 1-23

Cocolin L., Manzano M., Cantoni C. & Comi G. (2000) Development of a rapid method for the identification of Lactobacillus spp. isolated from naturally fermented Italian sausages by using PCR-temperature gradient gel electrophoresis (TGGE). *Letters in Applied Microbiology*, 30, 126-130

Colwell R.K. & Coddington J.A. (1994) Estimating Terrestrial Biodiversity through Extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 345, 101-118

Condit R., Ashton P., Bunyavejchewin S., Dattaraja H.S., Davies S., Esufali S., Ewango C., Foster R., Gunatilleke I.A.U.N., Gunatilleke C.V.S., Hall P., Harms K.E., Hart T., Hernandez C., Hubbell S., Itoh A., Kiratiprayoon S., LaFrankie J., de Lao S.L., Makana J.-R., Noor M.N.S., Kassim A.R., Russo S., Sukumar R., Samper C., Suresh H.S., Tan S., Thomas S., Valencia R., Vallejo M., Villa G. & Zillio T. (2006) The Importance of Demographic Niches to Tree Diversity. *Science*, 313, 98-101

Condit R., Pitman N., Leigh E.G., Chave J., Terborgh J., Foster R.B., Nunez P., Aguilar S., Valencia R., Villa G., Muller-Landau H.C., Losos E. & Hubbell S.P. (2002) Beta-diversity in tropical forest trees. *Science*, 295, 666-669

Corbet A.S. (1941) The distribution of butterflies in the Malay Peninsula. *Proceedings of the Royal Entomological Society of London (A)*, 16, 101–116

Curtis T.P. & Craine N.G. (1998) The comparison of the diversity of activated sludge plants. *Water Science and Technology*, 37, 71-78

Curtis T.P., Head I.M. & Graham D.W. (2003) Theoretical Ecology for engineering biology. *Environmental Science & Technology*, 37, 64A-70A

Curtis T.P., Head I.M., Lunn M., Sloan W.T., Schloss P.D. & Woodcock S. (2006) What is the Extent of Prokaryotic Diversity? *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 361, 2023-2037

Curtis T.P. & Sloan W.T. (2004) Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Current Opinion in Microbiology*, 7, 221-226

Curtis T.P. & Sloan W.T. (2005) Exploring Microbial Diversity - A Vast Below. *Science*, 309, 1331-1333

Curtis T.P., Sloan W.T. & Scannell J.W. (2002) Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 10494-10499

Daims H., K. S. & Wagner M. (2005) Fluorescence in situ hybridization for the detection of prokaryotes. In: *Advanced Methods in Molecular Microbial Ecology* (eds. Osbourne AM & Smith CJ), pp. 213-239. Bios-Garland, Abingdon

Davenport R.J., Curtis T.P., Goodfellow M., Stainsby F.M. & Bingley M. (2000) Quantitative use of fluorescent in situ hybridization to examine relationships between mycolic acid-containing actinomycetes and foaming in activated sludge plants. *Appl. Envir. Microbiol.*, 66, 1158-1166

de Candolle A. (1855) *Geographie botanique raisonnee: ou exposition des faites principaux et des lois concervant la distribution geographique des plantes de l'epoque actualle.* Maison, Paris.

DeGroot M.H. & Schervish M.J. (2001) *Probability and Statistics (3rd Edition).* Addison Wesley, Boston, MA.

Devereux R. & Willis S.G. (1995) Amplification of ribosomal RNA sequences. In: *Molecular Microbial Ecology Manual* (eds. Akkermans ADL, van Elsas JD & de Bruijn FJ), pp. 1-11. Kluwer, Dordrecht

Downing A.L., Painter H.A. & G. K. (1964) Nitrification in the Activated Sludge Process. *Journal of the Institute of Sewage Purification*, 63, 130-158

Dunbar J., Barns S.M., Ticknor L.O. & Kuske C.R. (2002) Empirical and Theoretical Bacterial Diversity in Four Arizona Soils. *Applied and Environmental Microbiology*, 68, 3035-3045

Enquist B.J., Sanderson J. & Weiser M.D. (2002) Modeling macroscopic patterns in ecology. *Science*, 295, 1835-1836

Fenchel T. (2003) Biogeography for bacteria. *Science*, 301, 925-926

Fenchel T. & Finlay B.J. (2005) Bacteria and Island Biogeography. *Science*, 309, 1997-1999

Fisher R.A., Corbet A.S. & Williams C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12, 42–58

Gans J., Wolinsky M. & Dunbar J. (2005) Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. *Science*, 309, 1387 - 1390

Giovannoni S.J. (1991) The polymerase chain reaction. In: *Nucleic Acid Techniques in Bacterial Systematics.* (eds. Stackebrandt E & Goodfellow M), p. 177–201. John Wiley & Sons,, New York

Gotelli N.J. & McGill B.J. (2006) Null Versus Neutral Models: What's The Difference? *Ecography*, 29, 793-800

Green J.L., Holmes A.J., Westoby M., Oliver I., Briscoe D., Dangerfield M., Gillings M. & Beattie A.J. (2004) Spatial scaling of microbial eukaryote diversity. *Nature*, 432

Grimmett G. & Stirzaker D. (2001) *Probability and Random Processes: Third Edition.* OUP, Oxford.

Grimmett G. & Welsh D. (1986) *Probability: an Introduction.* OUP, Oxford.

Hai F.I., Yamamoto K. & Fukushi K. (2006) Development of a submerged membrane fungi reactor for treatment of Textile Wastewater. *Desalination*, 192, 315-322

Harte J. (2003) Tail of death and resurrection. *Nature*, 424, 1006-1007

Harte J. (2004) The value of null theories in ecology. *Ecology*, 85, 1792-1794

Harte J., Kinzig A. & Green J. (1999) Self-similarity in the distribution and abundance of species. *Science*, 284, 334-336

He H.L. & Gaston K.J. (2003) Occupancy, spatial variance, and the abundance of species. *American Naturalist*, 162, 366-375

Head I.M., Saunders J.R. & Pickup R.W. (1998) Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecology*, 35, 1-21

Heuer H., Wieland G., Schonfeld J., Schnwalder A., Gomes N.C.M. & Smalla K. (2001) Bacterial community profiling using DGGE or TGGE analysis. In: *Environmental Molecular Microbiology: Protocols and Applications* (ed. Rochelle PA), p. 177–190. Horizon Scientific Press

Horner-Devine M.C., Lage M., Hughes J.B. & Bohannan B.J.M. (2004) A taxa-area relationship for bacteria. *Nature*, 432, 750-753

Hubbell S.P. (2001) *The unified neutral theory of biodiversity and biogeography.* Princeton University Press, Princeton.

Jaccard P. (1901) Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 241-272

Jagers P. & Klebaner F. (2003) Random variation and concentration effects in PCR. *Journal of Theoretical Biology*, 224, 299-304

James A. & Elliott D.J. (1984) Activated Sludge Models. In: *Water Quality Modelling* (ed. James A). Wiley, Chichester

Karlin S. & McGregor J. (1959) A Characterization of Birth and Death Processes. *Proceedings of the National Academy of Sciences of the United States of America*, 45, 375-379

Kassen R., Buckling A., Bell G. & Rainey P.B. (2000) Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature*, 406, 508-512

Kilburn P.D. (1966) Analysis of the species-area relation. *Ecology*, 47, 831-843

Kowalchuk G.A., de Bruijn F.J., Head I.M. & Akkermans A.D.L. (2005) *Molecular Microbial Ecology Manual.*

Laland K.N., Odling-Smee J. & Feldman M.W. (1999) Evolutionary consequences of niche construction and their implications for ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 10242-10247

Lane D.J. (1991) 16S/23S rRNA Sequencing. In: *Nucleic Acid Techniques in Bacterial Systematics* (eds. Stackebrandt E & Goodfellow M). Wiley, Chichester

Lavin M., Wojciechowski M.F., Richman A. & Sanderson M. (2001) Identifying Tertiary radiations of Fabaceae in the Greater Antilles: Alternatives to cladistic vicariance analysis. *International Journal of Plant Science*, 162, S53-S76

Lawrence A. & McCarty P.L. (1971) *Journal of the Sanitary Engineering Division - ASCE*, 97

Leclerc M., Delgènes J.-P. & Godon J.-J. (2004) Diversity of the archaeal community in 44 anaerobic digesters as determined by single strand conformation polymorphism analysis and 16S rDNA sequencing. *Environmental Microbiology*, 6, 809-819

Levin S.A. (2000) Multiple scales and the maintenance of biodiversity. *Ecosystems*, 3, 498-506

Linacre C. (2004) Diversity and the quantification of ammonia oxidising bacteria and denitrification from turbidity maximum of estuaries. In: *Department of Civil Engineering and Geosciences*. University of Newcastle Upon Tyne, Newcastle Upon Tyne

MacArthur R. (1960) On the relative abundance of species. *The American Naturalist*, 874, 25-36

MacArthur R. & Wilson E.O. (1967) *The Theory of Island Biogeography*. Princeton Univerity Press, Princeton, NJ.

Magurran A.E. (1996) *Ecological diversity and its measurement*. Princeton, Princeton, NJ.

Mangin I., Bonnet R., Seksik P., Rigottier-Gois L., Sutren M. & Bouhnik Y. (2004) Molecular inventory of faecal microflora in patients with Crohn's disease. *FEMS Microbiology Ecology*, 50, 25-36

May R.M. (1975) Patterns of species abundance and diversity. In: *Ecology and Evolution of Communities* (eds. Cody ML & Diamond JM), pp. 81-120. Harvard University Press, Harvard

McGill B.J. (2003) A test of the unified neutral theory of biodiversity. *Nature*, 422, 881-885

McGill B.J., Maurer B.A. & Weiser M.D. (2006) Empirical evaluation of Neutral Theory. *Ecology*, 87, 1411-1423

McKane A., Alonso D. & Sole R.V. (2004) Analytic solution of Hubbell's model of local community dynamics. *Theoretical Population Biology*, 65, 67–73

Metcalf L. & Eddy H.P. (1914) *American Sewage Practice vol. 1: Design of Sewers.* McGraw-Hill, New York, NY.

Moffett S., Brown D.A. & Under M.E. (2000) Lipid-dependent targeting of G proteins into rafts. *J. Biol. Chem.*, 275, 2191-2198

Mouquet N. & Loreau M. (2003) Community patterns in source-sink metacommunities. *American Naturalist*, 162, 544-557

Muneta Y., Shimoji Y., Yokomizo Y. & Mori Y. (1999) Moleculer Cloning of Porcine Interleuk in-1Beta Converting Enzyme and Differential Gene Expression of IL-1Beta Converting Enzyme, IL-1Beta and IL-18 in Porcine Alveolar Macrophages. *Journal of Interferon and Cytokine Research*, 19, 1289-1296

Muyzer G. (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. *Curr Opin Microbiol*, 2, 317–322

Muyzer G. & Smalla K. (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek*, 73, 127–141

Newman P. & Mouritz M. (1996) Principles and Planning opportunities for community scale systems of water and waste management. *Desalination*, 10, 339-354

Pielou E.C. (1966) The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13, 131-144

Pielou E.C. (1975) *Ecological Diversity.* Wiley, New York.

Pimm S.L. (1982) *Food Webs.* University of Chicago Press, Chicago, IL.

Porges N., Jasewicz L. & Hoover R.S. (1956) Principles of biological oxidation. In: *Biological treatment of sewage and industrial wastes* (eds. McCabe BJ & Eckenfelder WW), pp. 25-48. Reinhold, New York, NY

Preston F.W. (1948) The commonness and rarity of species. *Ecology*, 29, 254-283

Preston F.W. (1957) In. Pennsylvania State University

Preston F.W. (1962) The canonical distribution of commonness and rarity. *Ecology*, 43, 185-215

Price J., Droege S. & Price A. (1995) *The Summer Atlas of North American Birds.* Academic, San Diego.

Rittmann B.E., Hausner M., Löffler F., Love N.G., Muyzer G., Okabe S., Peccia J., Raskin L. & Wagner M. (2006) A Vista for Microbial

Ecology and Environmental Biotechnology: A consensus view for the partnership of microbial ecology and environmental biotechnology. *Environmental Science & Technology*, 40, 1096-1103

Rittmann B.E. & McCarty P.L. (2001) Environmental Biotechnology: Principles and Applications.

Robinson J.A. & Tiedje J.M. (1983) Nonlinear estimation of Monod growth kinetic parameters from a single substrate depletion curve. *Appl. Envir. Microbiol.*, 45, 1453-1458

Routledge R.D. (1980) The form of species abundance distributions. *Journal of Theoretical Biology*, 82, 547–58.

Rowan A.K., Snape J.R., Fearnside D., Barer M.R., Curtis T.C. & Head I.M. (2003) Composition and diversity of ammonia-oxidising bacterial communities in wastewater treatment reactors of differeny design treating identical wastewater. *FEMS Microbiol. Ecol.*, 43, 195-206

Saldinger M. (1992) *Small scale Wastewater Treatment Technologies - A GuideSmall scale Wastewater Treatment Technologies - A Guide.* Murdoch University, Perth, Western Australia.

Sambrook J., Fritsch E.F. & Maniatis T. (1989) *Molecular cloning: a laboratory manual.* Cold Spring Harbor Press, Cold Spring Harbor, NY.

Saunders D.A., Hobbs R.J. & Margules C.R. (1991) Biological consequences of ecosystem fragmentation: a review. *Conservation Biology*, 5, 18-32

Sawyer C.N. (1965) Milestones in the development of the activated sludge process. *Water Pollution Control Federation*, 62, 151-162

Scheiner S.M. (2003) Six types of species-area curves. *Global Ecology & Biogeography*, 12, 441-447

Schmidt L.D. (1998) *The Engineering of Chemical Reactions.* Oxford University Press, Oxford.

Shannon C.E. & Weaver W. (1949) *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, IL.

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

Sloan W.T., Lunn M., Woodcock S., Head I.M., Nee S. & Curtis T.P. (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, 8, 732-740

Sloan W.T., Woodcock S., Lunn M., Head I.M. & Curtis T.P. (2007) Modelling taxa-abundance distributions in microbial communities using environmental sequence data. *Microbial Ecology*, In Press

Sogin M.L., Morrison H.G., Huber J.A., Welch D.M., Huse S.M., Neal P.R., Arrieta J.M. & Herndl G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103, 12115-12120

Speece R.E. & McCarty P.L. (1964) Nutrient requirements and biological solids accumulation in anaerobic digestion. *Water Pollution Research*, 67, 294-301

Spellman F.R. (1997) *Microbiology for Water/Wastewater Operators.* Technomic Publishing Co. Inc., Lancaster, PA.

Sugihara G. (1980) Minimal community structure: an explanation of species abundance patterns. *The American Naturalist*, 116, 770-787

Thingstad T.F. & Lignell R. (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat.Microb.Ecol.*, 13, 19-27

Tilman D. (1976) Ecological competition between algae: Experimental confirmation of resource-based competition theory. *Science*, 192, 463-465

Tilman D. (1994) Competition and Biodiversity in Spatially Structured Habitats. *Ecology*, 75, 2-16

Tonn W.M. & Magnuson J.J. (1982) Patterns in the species composition and richness of fish assemblages in Northern Wisconsin lakes. *Ecology*, 63, 1149-1166

Torsvik V., Øvreås L. & Thingstad T.F. (2002) Prokaryotic Diversity—Magnitude, Dynamics, and Controlling Factors. *Science*, 296, 1064-1066

Vallade M. & Houchmandzadeh B. (2003) Analytic solution of a neutral model of biodiversity. *Physical Review E*, 68, 061902 1-5

Volkov I., Banavar J.R., Hubbell S.P. & Maritan A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035-1037

Volkov I., Banavar J.R. & Maritan A. (2006) Comment on "Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil". *Science*, 313, 918a

Volterra V. (1931) Variations and fluctuations of the number of individuals in animal species living together. In: *Animal Ecology*. McGraw-Hill, New York, NY von Wintzingerode F., Göbel U.B. & Stackebrandt E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.*, 21, 213-229

Wagner M. (2000) Phylogeny of all recognized species of ammonia oxidizers based on comparative 16S rRNA and amoA sequence analysis: Implications for molecular diversity surveys. *Applied and Environmental Microbiology*, 66, 5368-5382

Wagner M. & Loy A. (2002) Bacterial community composition and function in sewage treatment systems. *Curr. Opin. Biotechnol.*, 13, 218-227

Whitman W.B., Coleman D.C. & Wiebe W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 6578-6583

Williams C.B. (1964) *Patterns in the Balance of Nature and Related Problems of Quantitative Biology.* Academic Press, London.

Woodcock S., Curtis T.P., Head I.M., Lunn M. & Sloan W.T. (2006) Taxa-Area Relationships for Microbes: the Unsampled and the Unseen. *Ecology Letters*, 9, 805-812

Woodcock S., van der Gast C.J., Bell T., Lunn M., Curtis T.P., Head I.M. & Sloan W.T. (2007 In Press) Neutral assembly of bacterial communities. *FEMS Microbiol. Ecol.*

Wootton J.T. (2005) Field-parameterization and experimental test of the neutral theory of biodiversity. *Nature*, 433, 309-312

Zwart G., Crump B.C., Agterveld M., Hagen F. & Han S.K. (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology*, 28, 141-155