Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea Magistrale in Fisica

Ecological modelling for next generation sequencing data

Relatore:
Prof. Gastone Castellani

Correlatore:
Dott. Daniel Remondini

Presentata da:
Claudia Sala

Sessione II
Anno Accademico 2012/2013

Abstract

Le tecniche di next generation sequencing costituiscono un potente strumento per diverse applicazioni, soprattutto da quando i loro costi sono iniziati a calare e la qualità dei loro dati a migliorare.
Una delle applicazioni del sequencing è certamente la metagenomica, ovvero l'analisi di microorganismi entro un dato ambiente, come per esempio quello dell'intestino. In quest'ambito il sequencing ha permesso di campionare specie batteriche a cui non si riusciva ad accedere con le tradizionali tecniche di coltura.
Lo studio delle popolazioni batteriche intestinali è molto importante in quanto queste risultano alterate come effetto ma anche causa di numerose malattie, come quelle metaboliche (obesità, diabete di tipo 2, etc.).
In questo lavoro siamo partiti da dati di next generation sequencing del microbiota intestinale di 5 animali (16S rRNA sequencing) (Jeraldo et al.[35]). Abbiamo applicato algoritmi ottimizzati (UCLUST) per clusterizzare le sequenze generate in OTU (Operational Taxonomic Units), che corrispondono a cluster di specie batteriche ad un determinato livello tassonomico.
Abbiamo poi applicato la teoria ecologica a master equation sviluppata da Volkov et al.[49] per descrivere la distribuzione dell'abbondanza relativa delle specie (RSA) per i nostri campioni. La RSA è uno strumento ormai validato per lo studio della biodiversità dei sistemi ecologici e mostra una transizione da un andamento a logserie ad uno a lognormale passando da piccole comunità locali isolate a più grandi metacomunità costituite da più comunità locali che possono in qualche modo interagire.
Abbiamo mostrato come le OTU di popolazioni batteriche intestinali costituiscono un sistema ecologico che segue queste stesse regole se ottenuto usando diverse soglie di similarità nella procedura di clustering.
Ci aspettiamo quindi che questo risultato possa essere sfruttato per la comprensione della dinamica delle popolazioni batteriche e quindi di come queste variano in presenza di particolari malattie.

Contents

Introduction

Sequencing is the mean to determine the primary structure of a biopolymer, that is for example the exact order of nucleotides in a strand of DNA.
Nowadays, sequencing techniques are assuming an increasingly important role, particularly since their costs began to decline and their methods became more simple and widespread. These next generation sequencing techniques, which developed since the mid 1990s, are enabling us to gather many more times sequence data than was possible a few years ago.
Metagenomics is one of the many fields which exploit sequencing. In particular, with metagenomics we mean the collective genomes of microbes within a given environment, using indeed sequencing techniques to sequence particular strands of the genome of microorganisms. To study bacteria populations, for example, one sequences the 16S ribosomal RNA, which is a component of the 30S small subunit of prokaryotic ribosomes.
Ribosomal RNA is suitable for phylogenetic studies since it is a component of all self-replicating systems, it is readily isolated and its sequence changes but slowly with time, permitting the detection of relatedness among very distant species.
Metagenomics is one of the fastest advancing fields in biology [38]. By allowing access to the genomes of entire communities of bacteria, virus and fungi that are otherwise inaccessible, metagenomics is extending our comprehension of the diversity, ecology, evolution and functioning of the microbial world. The continuous and dynamic development of faster sequencing techniques, together with the advancement of methods and algorithms to cope with exponentially increasing amount of data generated are expanding our capacity to analyze microbial communities from an unlimited variety of habitats and environments.
In particular, exploiting next-generation sequencing techniques we became able to sample and study the gut microbiota biodiversity in order to understand how and why it results modified in many pathologies, such as in metabolic diseases and type 2 diabetes.
Patients affected by these pathologies, in fact, exhibit a certain degree of gut bacterial dysbiosis, that constitutes at the same time an effect but also a causal element of the pathology.

Many ecological theories have been proposed to understand and describe the biodiversity of ecological communities, a problem not completely solved by now. A common element of these theories is the idea that, to study the biodiversity of a system, we shall look at the relative species abundance distribution (RSA) rather than at the static coexistence among species, since ecological populations are evolving and not static systems [34]. Furthermore, it seems that these RSA distributions vary in many ways in different ecosystems, but always show somehow similar trends.
Volkov et al. [49] suggested a simple dynamical model which well describes RSA distribution of many different ecological systems, such as for example that of the coral-reef community, starting from the chemical master equation of a birth-death process. They thus obtained a negative binomial distribution with a shape parameter linked to immigration.
What we would like to do in this work is to exploit this dynamical model in order to describe data from gut microbiota populations, acquired through next generation sequencing techniques. For this purpose, we will described the functioning and applications of sequencing techniques (chapter 1) and the biomedical problems for which we are interested in analyzing the gut microbiota biodiversity (chapter 2). Thereafter, we will face the mathematical aspects of the chemical master equation (chapter 3), that we will exploit in the description of ecological models (chapter 4). In chapter 5 we will describe our dataset, which is a collection of 5 animals gut microbiota data from Jeraldo et al. [35], generated with next-generation sequencing. Then we will explain how we processed these data through specific optimized sequencing analysis algorithms, obtaining OTUs (clusters) which correspond to bacteria species. Finally we will show our results in the form of Preston plots, fitted with a gamma-like function, which corresponds to the continuous form of the negative binomial distribution predicted by Volkov et al. [49].

Chapter 1

Sequencing

In this chapter, first of all, we will explain the main sequencing applications, with a particular attention to that of gut microbiota, to give an idea of the importance of this technique. Then we will explore the most common sequencing methods, from Maxam-Gilbert sequencing to next-generation sequencing. Finally, we will give an insight of the principal algorithms that permit to analyze this kind of data, through alignment, clustering, distance computation and taxonomic assignment.

In general with the term 'sequencing' we refer to the means to determine the primary structure of a biopolymer. There are different types of sequencing (DNA sequencing, RNA sequencing, protein sequencing and ChIP sequencing) and different techniques to realize it. In particular we focus on RNA/DNA sequencing, that is the process of determining the precise order of nucleotides within a strand of RNA/DNA, i.e. of the four bases (adenine, guanine, cytosine, and thymine/uracil). For insights on DNA/RNA structure and the main biological processes exploited in sequencing, we refer to appendix B.

1.1 Sequencing Applications

Sequencing vs Microarray Nowadays, sequencing techniques are assuming an increasingly important role, particularly since their costs began to decline and their methods became more simple and widespread. Thus, researchers are choosing sequencing over the more common technique of microarrays, and not only for their genomic applications [22].
A DNA microarray (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface. These devices are used to measure the expression levels of large numbers of genes simultaneously

or to genotype multiple regions of a genome. Each DNA spot contains picomoles (10^{12} moles) of a specific DNA sequence, known as probes. These probes can be composed by short sections of genes or other DNA elements that are used to hybridize a cDNA or cRNA (where 'c' stands for 'complementary') sample (target) under high-stringency conditions. Probes are placed in known positions of the solid surface and are fluorophore-, silver-, or chemiluminescence- labeled, so that a probe-target hybridization can be detected and quantified. Since the positions and the sequences of the probes are known, we can determine the sequences present on our target, since they will hybridize with their complementary probe.
The main shortcoming of microarrays is that the probe matrix can include just a limited number of sequences, thus we have to previously know the probes that we need, that is which sequences we are going to analyze. So, for example *de novo* sequencing is not possible.
Next-generation sequencing methods provide a valid alternative to microarrays for some applications, such as chromatin immunoprecipitation, while for others, like cytogenetics, the transition between these two techniques has barely begun. The reasons for this are that, first of all, microarrays' longtime use as a genomics tool means many researchers are very comfortable using them, and that sample labeling, array handling and data analysis methods are tried and true. Secondly that, despite sequencing advancements, expression arrays are still cheaper and easier when processing large numbers of samples.
Nevertheless, the fast development of sequencing in producing high throughput data at lowering costs makes us suppose that these techniques will replace microarrays also in the fields in which these still rule.

Applications fields We now give a short overview of the main fields of sequencing applications reported in [20].
As already mentioned, one of the main applications of sequencing, is that of ***de novo* sequencing**. The aim here is to sequence a genome not yet known.
A second field is the sequencing of the **trascriptome** (RNAseq) and of **microRNA** (a small non-coding RNA molecule of about 22 nucleotides, which functions in transcriptional and post-transcriptional regulation of gene expression). This can be considered another great tool to analyze the biological functions inside a cell, since it gives informations about the gene expression in different tissues or at different conditions inside a certain tissue, that can be exploited in studies of RNA interference or more in general in epigenetic studies.
A third field of application is that of **resequencing**, where an whole already sequenced genome needs to be resequenced, for example to identify any genetic deficiency such as mutations, insertions, deletions or alterations in the number of

a gene copies. This kind of analysis is very useful to understand the development of some pathologies and to find eventual clinical treatments.
Another important field of application, in which our work is included, is the sequencing of microbiotic communities for **metagenomic** studies. Metagenomic sequencing allows to analyze samples directly taken from the microbiotic environment, avoiding the issue of growing bacteria artificially. In fact, as we will underline in section 2.1, traditional clonal culture techniques result biased and cannot access the vast majority of organisms within a community. In this contest the sequenced strands are those of the 16S ribosomal RNA (see appendix C), since this results highly conserved between different species of bacteria and archaea and thus can be used for phylogenetic and biodiversity studies.
One of the widest metagenomic studies, was that begun in 2003 by Craig Venter, leader of the privately funded parallel of the Human Genome Project, who has led the Global Ocean Sampling Expedition (GOS), circumnavigating the globe and collecting metagenomic samples throughout the journey. All of these samples were sequenced using shotgun sequencing, in the hope that new genomes (and therefore new organisms) would be identified. The pilot project, conducted in the Sargasso Sea, found DNA from nearly 2000 different species, including 148 types of bacteria never before seen [48]. Analysis of the metagenomic data collected during Venter's journey also revealed two groups of organisms, one composed by taxa adapted to environmental conditions of 'feast or famine', and a second composed by relatively fewer but more abundantly and widely distributed taxa primarily composed by plankton. Thus, this study resulted in an important turning point in ocean biodiversity knowledge, and above all it showed the great potentialities of nowadays sequencing techniques.
Finally, the last application that we are going to mention is that of **Chromatin ImmunoPrecipitation-Sequencing (ChIP-seq)**. This technique is used to study the interaction between DNA and regulatory proteins. In fact, immunoprecipitation allows to identify the positions on the DNA strand, on which transcriptional factors, histones or proteins can be bound to control DNA replication. Thus, with different sequencing we can understand the influence of environmental alterations on the phenotype.

1.2 DNA Sequencing techniques

After explaining the great potentiality of sequencing, let us now give an insight of how this techniques work.
Sequencing techniques, are able to analyze just fragments of DNA. The fragment of DNA that is being read is called 'read', and it is composed by at most a hundred

of nucleotides. These reads then need to be processed and assembled to build the unknown sequence. We will see later the most common algorithms to analyze these reads, but for now let us describe the main techniques of DNA sequencing.

First generation DNA sequencing methods

Maxam-Gilbert sequencing Maxam-Gilbert sequencing was the first widely-adopted method for DNA sequencing, developed by Allan Maxam and Walter Gilbert in 1976-1977 and also known as the chemical degradation method. This technique is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides [41].
The first step in Maxam-Gilbert sequencing is a radioactive labeling at one 5'-end of the DNA fragment that we want to sequence. This is typically done through a kinase reaction using $\mathrm{gamma}^{32}\mathrm{P}$ ATP, where $^{32}\mathrm{P}$ is a radioactive isotope of phosphorus that decays into sulfur-32 by beta decay with a half-life of about 14 days.
The DNA fragment is then denaturated, i.e. its double strand is separated into two single strands through the breaking of the hydrogen bonds between them. After this procedure the DNA fragment is subjected to four specific chemical reactions, that generate breaks of different sizes and in different positions:

- dimethyl sulfate (DMS) plus piperdine cleaves at G;
- DMS plus piperdine and formic acid cleaves at A or G;
- hydrazine plus piperdine cleavs at C or T;
- hydrazine in a saline solution (NaCl) plus piperdine cleaves at C.

As shown in figure 1.1, reaction products are then electrophoresed on a polyacrylamide denaturing gel for size separation. To visualize the fragments, the gel is exposed to an X-ray film for autoradiography, yielding a series of dark bands each showing the location of identical radiolabeled DNA molecules. The sequence can be deduced from the presence or the absence of certain fragments.

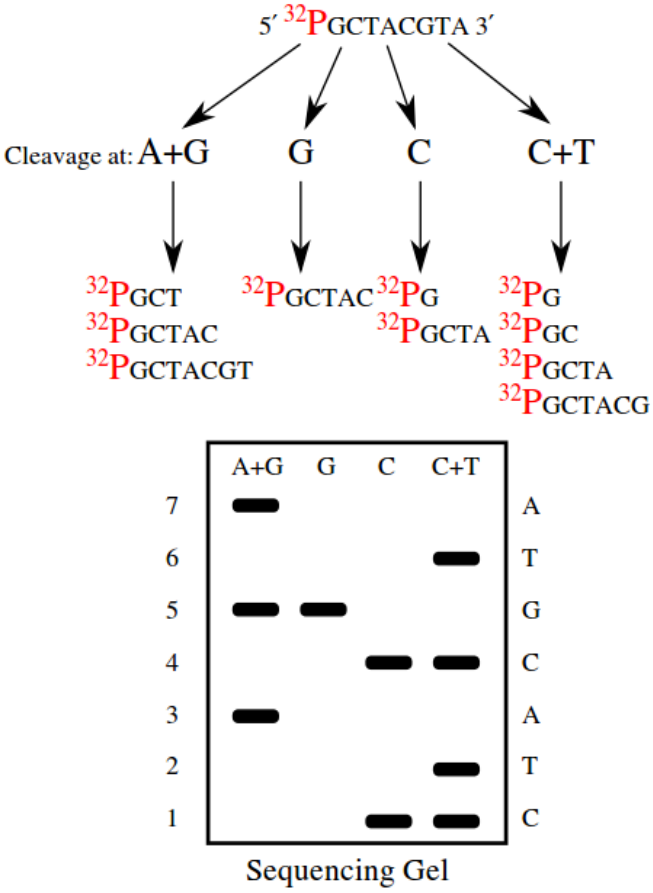


Figure 1.1: An example Maxam-Gilbert sequencing reaction from [6]. In level 1 there will be the shortest sequences and in level 7 the longest.

Sanger sequencing Sanger sequencing, also called chain-terminator sequencing, has been developed by Fredrick Sanger and colleagues in 1977 and was the most widley-used method for about 25 years.
Similarly to the Maxam-Gilbert method, the DNA sample is subjected to four separate sequencing reactions. Each reaction contains the DNA polymerase, plus three of the four standard deoxynucleotides, that are the DNA nucleosides triphosphate (dATP, dGTP, dCTP and dTTP) required for the DNA extension and only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP), which are modified nucleotides that terminate the DNA strand elongation, since they lack the 3'-OH group required for the formation of the phosphodiester with the following nucleotide. When the specific ddNTP included in the reaction is incorporated in the DNA strand, the polymerase ceases the extension of DNA and we obtain DNA

fragments that terminate with a specific nucleotide. The resulting DNA fragments are then denatured and separated by size using gel electrophoresis, like in the Maxam Gilbert method. To visualize the DNA bands automatically, the ddNTPs are also radioactively or fluorescently labeled, so that the DNA bands can be visualized by autoradiography or UV light and the DNA sequence can be directly read off the X-ray film or gel image [5].

The greatest limitations of Maxam - Gilbert and Sanger sequencing methods are that they are quite expensive and that they can be used just for fairly short strands (100 to 1000 basepairs). However, some improvements have been developed in order to allow the sequencing of longer strands.

Longer strands sequencing For longer targets such as chromosomes, common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA may then be cloned into a DNA vector and amplified in a bacterial host such as Escherichia coli. Amplification is required to have more copies of the same DNA fragment, in order to have more robust statistical informations. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence. There are two main methods used for this purpose: primer walking and shotgun sequencing.
Primer walking starts from the beginning of the DNA target, using the first 20 bases as primers for a PCR (polymerase chain reaction), that is a technique of DNA amplification, amplifying about 1000 bases and then sequencing them using the chain termination method. Then the method ‘walks’ on the DNA strand and uses the last 20 bases of the previous, now known, sequence as primers, and so on.
In shotgun sequencing, instead, DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain reads. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.

Next-generation DNA sequencing methods

The high demand for low-cost sequencing has driven the development of high-throughput sequencing (or next-generation sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently. Nowadays, these new methods can analyze up to 600 billions of bases (GB) in a ten days cycle, versus the one million of the first generation methods, and this of course coincide with a lowering in the costs. We can give an idea of the magnitudes we deal with considering that the HIV virus genome is 3.4 kB, the E. coli bacteria genome is 4.6 MB, and the human one is 3.2 GB [20]
The main shortcoming of these new technologies is the small length of the reads, that generates errors during the assembly phase, that should be considered in further analysis.
Let us describe three of the main next-generation sequencing techniques: 454 pyrosequencing, Illumina (Solexa) sequencing and SOLiD sequencing.

454 pyrosequencing Pyrosequencing is a method of DNA sequencing based on the 'sequencing by synthesis' principle; it was the first next-generation method available on the market (produced by 454 Life Sciences since 2005 and now owned by Roche Diagnostics). It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate released by nucleotide incorporation, rather than chain termination with dideoxynucleotides.
'Sequencing by synthesis' involves taking a single strand of the DNA sample that one wants to sequence and then synthesizing its complementary strand enzymatically. The pyrosequencing method is based on detecting the activity of DNA polymerase with another chemiluminescent enzyme. The template DNA is immobilized, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction, so that we can detect which base was actually added by the DNA polymerase at each step. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.
A parallelized version of pyrosequencing was developed by 454 Life Sciences [17], see fig.1.2. 454 Sequencing uses a large-scale parallel pyrosequencing system capable of sequencing roughly 400-600 megabases of DNA per 10-hour run.

In 454 pyrosequencing, DNA samples are first fractionated into smaller fragments (300-800 base pairs) and polished (made blunt at each end). Short adaptors, that are short, chemically synthesized, double stranded DNA molecules, are then ligated onto the ends of the fragments. These adaptors provide priming sequences for

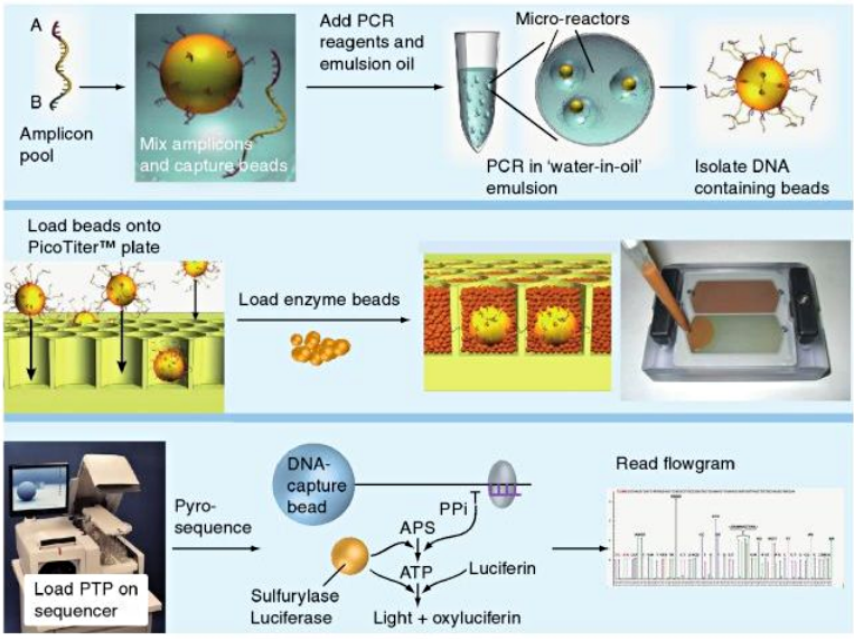


Figure 1.2: 454 Life Sciences Sequencing Technology. Pooled amplicons are clonally amplified in droplet emulsions. Isolated DNA-carrying beads are loaded into individual wells on a PicoTiter plate and surrounded by enzyme beads. Nucleotides are flowed one at a time over the plate and template-dependent incorporation releases pyrophosphate, which is converted to light through an enzymatic process. The light signals, which are proportional to the number of incorporated nucleotides in a given flow, are represented in flowgrams that are analyzed and a nucleotide sequence is determined for each read. Figure from [31].

both amplification and sequencing of the sample fragments. One adaptor (Adaptor B) contains a 5'-biotin tag for immobilization of the DNA fragments onto streptavidin-coated beads. Then, the non-biotinylated strand is released and used as a single-stranded template DNA (sstDNA).
Let us observe that there should be a great number of beads (around one million), so that each bead will carry just a single sstDNA molecule. The beads are then emulsified with the amplification reagents in a water-in-oil mixture. This leads to the formation of drops of water (containing the beads) in the oil mixture, where PCR amplification occurs. This is useful, since the amplification can be done in vitro, keeping the different fragments reactions separated. This part of the process results in bead-immobilized, clonally amplified DNA fragments.
Subsequently, the beads are placed onto a PicoTiterPlate device, that is composed of around 1.6 million wells, small enough to contain just one bead ($\sim 28\mu m$ of di-

ameter). The device is centrifuged to deposit the beads into the wells and the DNA polymerase is added, with also other smaller beads (containing two enzymes: sulfurylase and luciferase), which ensure that the DNA beads remain positioned in the wells during the sequencing reaction.
At this point also the sequencing reagents required by pyrosequencing are delivered across the wells of the plate. These include ATP sulfurylase, luciferase, apyrase, the substrates adenosine 5 phosphosulfate (APS) and luciferin and the four deoxynucleoside triphosphates (dNTPs). The four dNTPs are added sequentially in a fixed order across the PicoTiterPlate device during a sequencing run. During the nucleotide flow, millions of copies of DNA bound to each bead are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Addition of one (or more) nucleotide(s) generates a light signal that is recorded by the CCD camera in the instrument and that is proportional to the number of nucleotides. This can be explained following the biochemical reaction that occurs when the dNTP is complementary to the next nucleotide on the fragment. In this case, the bound between the two bases will release pyrophosphate (PPi) in stoichiometrical amounts. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5 phosphosulfate. This ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP, that is proportional to the number of nucleotides bound by this dNTP type. Unincorporated nucleotides and ATP are then degraded by the apyrase enzyme, and the reaction can restart with another nucleotide.
Let us note that when the polymerase meets homopolymers, that are sequences of the same kind of nucleotide (e.g. AAAA), the contiguous bases are incorporated during the same cycle and their number can be deduced just through the intensity of the emitted light, that sometimes can be misleading.
Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, that is shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA.
Bisides the rapid evolution of 454 pyrosequencing technology for what concern the sequencing time (it allows to sequence one million fragments in 10 hours) and costs (even if it remains the most expensive technique of next-generation sequencing), this progresses have not been accompanied by a reassessment of the quality and accuracy of the sequences obtained. The mean error rate for this technology is in fact of 1.07% [33]. More importantly, this error rate is not randomly distributed; it occasionally rose to more than 50% in certain positions, and its distribution was linked to several experimental variables like the presence of homopolymers, the

position in the sequence, the size of the sequence and its spatial localization in PicoTiter plates.

Illumina (Solexa) sequencing Illumina (Solexa) sequencing was first commercialized by Solexa in 2006, a company later acquired by Illumina. This sequencing method is based on the reversible chain termination method described previously, and its general functioning can be subdivided in three phases: library generation, cluster preparation and sequencing.
During the first phase, DNA is fragmented and oligo-adaptors are added at each fragment sides for the next amplification process. Amplification occurs on flow-cell, that is a plate on which DNA molecules are attached and on which two different types of oligonucleotides are present. DNA fragments ends bind to these oligonucleotides, each end binding with its complementary nucleotide, so that a bridge structure is created, as shown in figure 1.3.

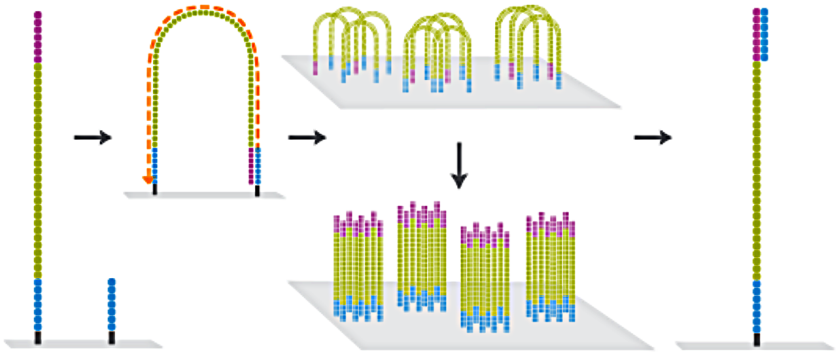


Figure 1.3: DNA ligated with adaptors is attached to the flow-cell; bridge amplification is performed; clusters are generated; the sequencing primers are synthetizes. Figure from [8].

At this point DNA polymerase synthesizes the complementary strands of our fragments, that are then denaturated. The hydrogen bonds are broken and we obtain again two separated strand, doubled compared to the beginning. The process is repeated to obtain a cluster of thousands of fragments, that however contain both the original strand and the complementary one. Thus, it is necessary to remove the antisense strands, before sequencing the samples.
In the last step, primers of the fragments of each cluster are synthesized. These primers are those sequences that start the sequencing reaction. So, sequencing can

be done on millions of clusters in parallel.

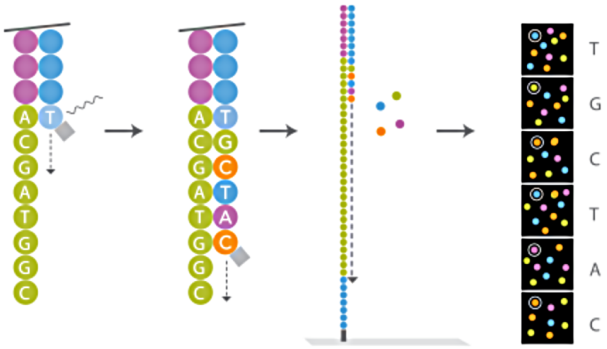


Figure 1.4: The first base is extended, read and deblocked; the above step is repeated on the whole strand; the fluorescent signals are read. Figure from [8].

Each step of the sequencing process involves a DNA polymerase and the four modified dNTP, which contain also a fluorescent marker and a reversible terminator. Markers of the four nucleotides react in different manner when subjected to a laser wave and this allows the identification of the sequenced base (see fig.1.4).
After each incorporation, a laser excites the fluorescent marker generating a light emission that allows the identification of the base. Then the terminator and the fluorescent label are removed, so that the next base can be sequenced.
With a single cycle, Illumina sequencer can read up to 6 billion reads in few days, with a number of bases ranging from 50 to 200, that means a total of almost 1000 GBases.

SOLiD sequencing Applied Biosystems' (now a Life Technologies brand) SOLiD technology employs sequencing by ligation. Like in 454 pyrosequencing, DNA fragments are bound to adaptors in order to immobilize them onto beads and to amplify them through emPCR. After denaturation, beads are placed on a glass support; the difference between this support and the PicoTiter plate is that in SOLiD there are no wells, thus the only limitation on the number of beads is due to their diameter, that now is much smaller than for the 454 technologies ($< 1\mu m$). In SOLiD, sequencing by synthesis is driven by DNA ligase, rather than polymerase, that is an enzyme that facilitates the joining of DNA strands together by catalyzing the formation of a phosphodiester bond, hence the acronym SOLiD (Sequencing by Oligonucleotide Ligation and Detection).

Each sequencing cycle needs a bead, a degenerate primer (which can bind all the four bases), a ligase and four dNTP 8-mer probe, which are eight bases in length with a free hydroxyl group at the 3' end, with a fluorescent dye at the 5' end and with a cleavage site between the fifth and sixth nucleotide. The first two bases (starting at the 3' end) are complementary to the nucleotides being sequenced, while bases 3 through 5 are degenerate and able to pair with any nucleotides on the template sequence.
First of all the primer hybridizes with the adaptor sequence, then the ligase allows the bound of a probe, followed by fluorescent emission from the dye; finally, the last three bases (6-7-8) of the 8-mer bound are removed together with the dye, to allow the analysis of subsequent bases.
Each couple of bases is associated with a particular color to allow the identification; however the labeling is not univocal, since we have 4 colors and 16 possible couples of bases. So, we can wonder why associating a color to each couple rather than to each base. Actually, the method used here is more convenient since using a 1-1 corrispondence is more probable to generate sequencing errors, while this method can help in avoiding them (see fig.1.5). Let us note that, since in each cycle we will sequence 2 nucleotides every 5, we will have to repeat the sequencing cycle five times to univocally determine each base (see fig.1.6).
During this process, just few 8-mers can be bound together (7, or at most 10), and this lead to very short reads (35-50 bases), but at the same time this procedure allows a minimization of the errors during each read scanning.

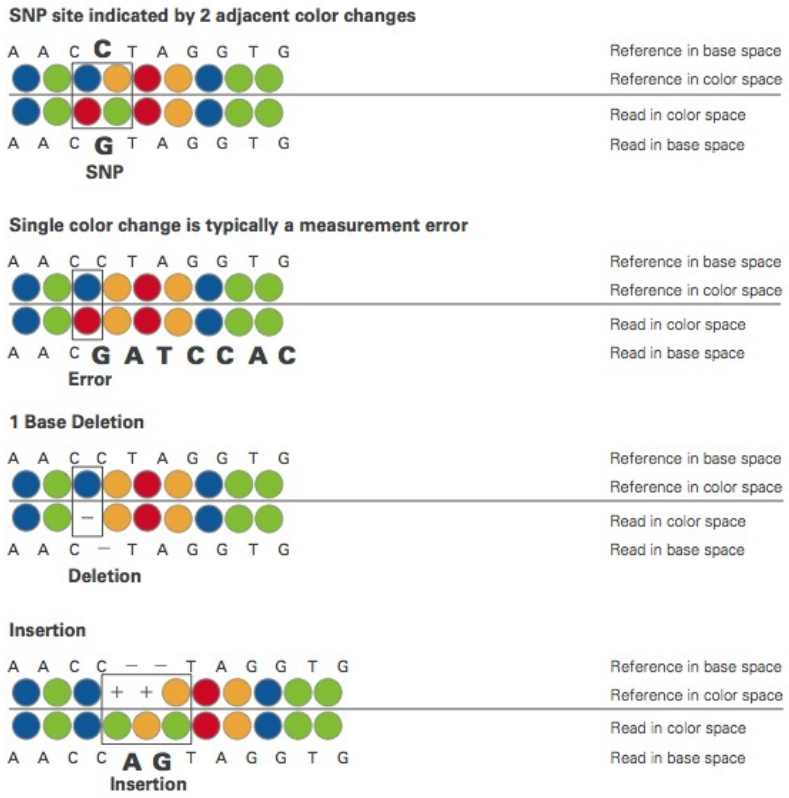


Figure 1.5: How SOLiD responds to single mutations, measurement errors, deletions and insertions. Figure from [9].

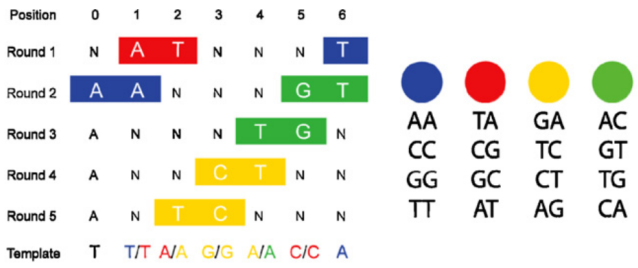


Figure 1.6: Color output of the template TAGACA. Because of the end-labeling, all sequences start with a T, therefore since the first signal output is blue, the first base-pair of the sequence must be a T, since the only blue probe that begins with an A (complementary to the T) is the AA probe. The machine uses the same logic to compute the entire strand. Figure from [39].

Specifications summary

Technology	Sanger Sequencing	Next Generation Sequencing			
Manufacturer	Applied Biosystems	Roche 454	Illumina	Life Technologies	
Model	ABI 3730XL	GS FLX Titanium XL+	HiSeq 2000 dual flow cell	SOLiD 4 System	Ion PGM
Bases per RUN	~ 96 Kb	700 Mb	600 Gb	100 Gb	1 Gb
Time per RUN	2 h	~1 day	~11 days	~14 days	4.5 h
Reads per RUN	96	1 million	6 billions (paired-end)	1.4 billions	5 millions
Reads length	up to 1000 bp	up to 1000 bp (mode 700 bp)	2*100 bp	2*50 bp	up to 400 bp

Figure 1.7: Specifications of different sequencing techniques. Figure from [20].

Computational requirements

The high performances of these next-generation techniques led to the need of also high computational ability for what concern both data storage and elaboration [20].
We have to consider, in fact, that for each sequenced base, we can have up to 16 byte (but also more), and that a Illumina or SOLiD run can need some Tbyte of memory, neglecting eventual backup and redundances. To give an idea of the amount of data produced by next-generation sequencing platform, we can refer to the 9 petabyte ($18 \cdot 10^{50}$ byte) generated in 2010 by the Sanger Insitute alone, that is one of the biggest sequencing center in the world.
Furthermore, these big amount of data need to be processed and analyzed: reads need to be assembled and/or aligned. Thus, besides the storage memory, also high-performances CPU and algorithms are needed.

1.3 Algorithms

Usually, sequencing data analysis includes the following processing procedures: alignment, distances computations, clustering and taxonomic assignment. Let us now describe the principal optimized algorithms to compute these elaborations. We will exploit these algorithms through QIIME (Quantitative Insights Into Microbial Ecology) [12], that is an open source software package for comparison and analysis of microbial communities, primarily based on high-throughput sequencing data generated on a variety of platforms, but also supporting analysis of other types of data (such as shotgun metagenomic data). QIIME takes users from their raw sequencing output through initial analyses such as OTU picking, taxonomic assignment, and construction of phylogenetic trees from representative sequences of OTUs, and through downstream statistical analysis, visualization, and production of publication-quality graphics.

1.3.1 Sequence alignment

Computational algorithms to sequence alignment generally fall into two categories: global alignments and local alignments.
Calculating a global alignment is a form of global optimization that 'forces' the alignment to span the entire length of all query sequences. These methods are more useful when the sequences in the query set are similar and of roughly equal size.
By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Thus, these methods are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying the regions of similarity. One motivation for local alignment is the difficulty of obtaining correct alignments in regions of low similarity between distantly related biological sequences, because mutations have added too much 'noise' over evolutionary time to allow for a meaningful comparison of those regions. Local alignment avoids such regions altogether and focuses on those with an evolutionary conserved signal of similarity.
There exist also hybrid methods, which attempt to find the best possible alignment that includes the start and end of one or the other sequence. This can be especially useful when the downstream part of one sequence overlaps with the upstream part of the other sequence. In this case, neither global nor local alignment would entirely work.
A variety of computational algorithms have been applied to the sequence align-

ment problem. These include slow but formally correct methods like dynamic programming, but also include efficient, heuristic algorithms or probabilistic methods designed for large-scale database search, that do not guarantee to find best matches.

Smith-Waterman Early alignment programs, such as the Smith-Waterman algorithm and the Needleman-Wunsch, which is a variation of the first one, used dynamic programming algorithms, that is methods based on the idea that to solve complex problems one can break them down into simpler subproblems. Often when using a more naive method, many of the subproblems are generated and solved many times. The dynamic programming approach seeks to solve each subproblem only once, thus reducing the number of computations: once the solution to a given subproblem has been computed, it is stored: the next time the same solution is needed, it is simply looked up. This approach is especially useful when the number of repeating subproblems grows exponentially as a function of the size of the input.
Dynamic programming algorithms are used for optimization (for example, finding the shortest path between two points, or the fastest way to multiply many matrices). A dynamic programming algorithm will examine all possible ways to solve the problem and will pick the best solution. Therefore, we can roughly think of dynamic programming as an intelligent, brute-force method that enables us to go through all possible solutions to pick the best one. If the scope of the problem is such that going through all possible solutions is possible and fast enough, dynamic programming guarantees finding the optimal solution.
The Smith-Waterman algorithm is a dynamic programming method which performs local sequence alignment with the guarantee of finding the optimal alignment [15].
The algorithm first builds a matrix H as follows:

$$\begin{aligned} H(i,0) &= 0; \text{ for } 0 \leq i \leq m \\ H(0,j) &= 0; \text{ for } 0 \leq j \leq n \end{aligned} \tag{1.1}$$

Then, if $a_i = b_j$ then $w(a_i, b_j) = w$(match) or if $a_i \neq b_j$ then $w(a_i, b_j) = w$(mismatch), thus for $1 \leq i \leq m, 1 \leq j \leq n$, we have

$$H(i,j) = \max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + w(a_i, b_j) & \text{Match/Mismatch} \\ H(i-1,j) + w(a_i, -) & \text{Deletion} \\ H(i,j-1) + w(-, b_j) & \text{Insertion} \end{array} \right\}. \tag{1.2}$$

where:

- a, b = strings that we want to align;
- m = length(a);
- n = length(b).

Let us now show an example from [15].
Sequence 1 = ACACACTA
Sequence 2 = AGCACACA
$w(\text{match}) = +2$
$w(a, -) = w(-, b) = w(\text{mismatch}) = -1$

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \mathbf{2} & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & \mathbf{3} & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & \mathbf{5} & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & \mathbf{7} & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & \mathbf{9} & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & \mathbf{11} & \mathbf{10} & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & \mathbf{12} \end{pmatrix} \tag{1.3}$$

$$T = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & \mathbf{0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ G & 0 & \mathbf{\uparrow} & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \mathbf{\nwarrow} & \nwarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \leftarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \mathbf{\nwarrow} & \leftarrow & \leftarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \mathbf{\nwarrow} \end{pmatrix} \tag{1.4}$$

To obtain the optimum local alignment, we start with the highest value in the matrix (i, j). Then, we go backwards to one of positions $(i - 1, j)$, $(i, j - 1)$, and $(i-1, j-1)$ depending on the direction of movement used to construct the matrix. We keep the process until we reach a matrix cell with zero value.
In the example, the highest value corresponds to the cell in position $(8, 8)$. The walk back corresponds to $(8, 8)$, $(7, 7)$, $(7, 6)$, $(6, 5)$, $(5, 4)$, $(4, 3)$, $(3, 2)$, $(2, 1)$, $(1, 1)$, and $(0, 0)$.
Once we've finished, we reconstruct the alignment as follows: starting with the last value, we reach (i, j) using the previously calculated path. A diagonal jump

implies there is an alignment (either a match or a mismatch). A top-down jump implies there is a deletion. A left-right jump implies there is an insertion.
For our example, we get:
Sequence 1 = A-CACACTA
Sequence 2 = AGCACAC-A
The Smith-Waterman algorithm is fairly demanding of time: to align two sequences of lengths m and n, $O(mn)$ time is required. Other algorithms such as BLAST, that we are now going to describe, reduce the amount of time required by identifying conserved regions using rapid lookup strategies, at the cost of exactness.

BLAST and FASTA BLAST (Basic Local Alignment Search Tool) and FASTA (FAST All) are heuristic algorithms, and as such they are designed for solving a problem more quickly when classic dynamic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. By trading optimality, completeness, accuracy, and/or precision for speed, a heuristic method can quickly produce a solution that is good enough for solving the problem at hand, as opposed to finding all exact solutions in a prohibitively long time. Thus heuristic algorithms are more practical for the analysis of the huge genome databases currently available.
BLAST is more time-efficient than FASTA by searching only for the more significant patterns in the sequences, yet with comparative sensitivity; thus we will focus mostly on BLAST.
To run, BLAST requires a query sequence to search for, and a sequence to search against (also called the target sequence) or a sequence database containing many target sequences. BLAST will find sub-sequences in the database which are similar to subsequences in the query. In typical usage, the query sequence is much smaller than the database, e.g., the query may be one thousand nucleotides while the database is several billion nucleotides.
The main idea of BLAST is that there are often high-scoring segment pairs (HSP) contained in a statistically significant alignment. BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm, that, as we already observed, is too slow for searching large genomic databases such as GenBank.
Let us report how BLAST basically works, as described in [18].

1. Remove low-complexity regions or sequence repeats in the query sequence, where 'low-complexity' region means a region of a sequence composed of

few kinds of elements. These regions might give high scores that confuse the program to find the actual significant sequences in the database, so they should be filtered out. The regions will be marked with an X (protein sequences) or N (nucleic acid sequences) and then be ignored by the BLAST program.

2. Make a list of all the k-letter words inside the query sequence, where k usually is 3 for proteins and 11 for nucleotides.

3. List the possible matching words. This step is one of the main differences between BLAST and FASTA. FASTA cares about all of the common words in the database and query sequences that are listed in step 2; however, BLAST only cares about the high-scoring words. The scores are created by comparing the words in the step 2 list with all the k-letter words and giving a score according to how many matching and non-matching words are present.

4. Organize the remaining high-scoring words into an efficient search tree. This allows the program to rapidly compare the high-scoring words to the database sequences.

5. Repeat step 3 to 4 for each k-letter word in the query sequence.

6. The BLAST program scans the database sequences for the high-scoring word of each position. If an exact match is found, this match is used to seed a possible un-gapped alignment between the query and database sequences.

7. Extend the exact matches to high-scoring segment pair (HSP). The original version of BLAST stretches a longer alignment between the query and the database sequence in the left and right directions, from the position where the exact match occurred. The extension does not stop until the accumulated total score of the HSP begins to decrease. A simplified example is presented in fig.1.8.

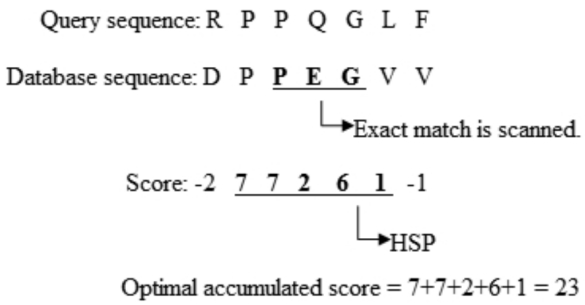


Figure 1.8: The process to extend the exact match. Figure from [18].

To save more time, a newer version of BLAST, called BLAST2, adopts a lower neighborhood word score threshold to maintain the same level of sensitivity for detecting sequence similarity. Therefore, the possible matching words list in step 3 becomes longer. Next, the exact matched regions, within distance A from each other on the same diagonal in fig.1.9, will be joined as a longer new region.
Finally, the new regions are then extended by the same method as in the original version of BLAST, and the HSPs' scores of the extended regions are then created as before.

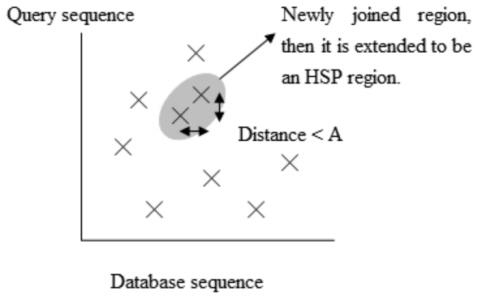


Figure 1.9: The positions of the exact matches. Figure from [18].

8. List all of the HSPs in the database whose score is higher then an empirically determined cutoff score S. By examining the distribution of the alignment scores modeled by comparing random sequences, a cutoff score S can be determined such that its value is large enough to guarantee the significance of the remaining HSPs.

9. Evaluate the significance of the HSP score (E-value), that is the number of times a random database sequence would give a score higher than S by chance.

10. Make two or more HSP regions into a longer alignment. Sometimes, we find two or more HSP regions in one database sequence that can be made into a longer alignment. This provides additional evidence of the relation between the query and database sequence. There are two methods, the Poisson method and the sum-of-scores method, to compare the significance of the newly combined HSP regions. Suppose that there are two combined HSP regions with the pairs of scores $(65, 40)$ and $(52, 45)$, respectively. The Poisson method gives more significance to the set with the maximal lower score ($45 > 40$). However, the sum-of-scores method prefers the first set, because $65 + 40$ (105) is greater than $52 + 45$ (97). The original BLAST uses the Poisson method; BLAST2 uses the sum-of scores method.

11. Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences. The original BLAST only generates un-gapped alignments including the initially found HSPs individually, even when there is more than one HSP found in one database sequence. BLAST2 produces a single alignment with gaps that can include all of the initially-found HSP regions. Note that the computation of the score and its corresponding E score is involved with the adequate gap penalties.

12. Report every match whose expect score is lower than a threshold parameter E.

Clustal W There are three main steps [42]:

1. all pairs of sequences are aligned separately in order to calculate a distance matrix giving the divergence of each pair of sequences;

2. a guide tree (or a user-defined tree) is calculated from the distance matrix;

3. the sequences are progressively pairwise aligned according to the branching order in the guide tree. Thus, first are considered the nearest sequences and then the farther. At each stage, gaps can be introduced.

In the original CLUSTAL programs, the pairwise distances are calculated giving a score to the number of matches and a penalty for each gap. The latest versions allow to choose between this method and the slower but more accurate scores from full dynamic programming alignments using two gap penalties, which differ if there is an opening gap or an extending one. These scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gaps excluded). Both of these scores are initially calculated as per cent identity scores and are converted to distances by dividing per 100 and subtracting from 1.0 to give the number of differences per site.
The main advantage of CLUSTALW on previous methods is that it gives a better quality without affecting the costs.

MUSCLE MUSCLE [28] is often used as a replacement for Clustal, since it typically (but not always) gives better sequence alignments and is significantly faster than Clustal, especially for larger alignments. However, it remains quite slow compared to other methods like NAST [4].
The main steps of MUSCLE are the same of CLUSTAL: distance matrix computation, guide tree computation, pairwise alignment following the guide tree. MUSCLE exploits the Kimura distance, which is a more accurate measurement even if it requires a previous alignment, and the subdivision of the tree in subtrees in which the profile of multiple alignment is computed so that, with a re-alignment of these profiles one can try to find an eventual better score.

UCLUST UCLUST [2] creates multiple alignments of clusters. Thus, it requires a first step of clustering, then a conversion to .fasta and finally the insertion of additional gaps. We will explain in more detail the clustering step in subsection 1.3.2.

NAST In NAST [27], an unaligned sequence is termed the ‘candidate’ and is matched to templates by comparison of 7-mers in common.
At first, a BLAST pairwise alignment is performed between the candidate and the template. As a result of the pairwise alignment performed by BLAST, new alignment gaps (hyphens) are introduced between the bases of the template whenever the candidate contains additional internal bases (insertions) compared with the template (fig.1.10 A, B). Any pairwise alignment algorithm must do this to compensate for nucleotides not shared by both sequences. This expansion, when intercalated with the original template spacing, results in candidates occupying more columns (characters) than the original template format (fig.1.10 C). Since

a consistent column count may be an option chosen by the user, the candidate-template alignment is compressed back to the initial number of characters with NAST. After insertion bases are identified (fig.1.10 C), a bidirectional search for the nearest alignment space (hyphen) relative to the insertion results in character deletion of the proximal place holders. Ultimately, local misalignments, spanning from the insertion base to the deleted alignment space, are permitted to preserve the global multiple sequence alignment format.

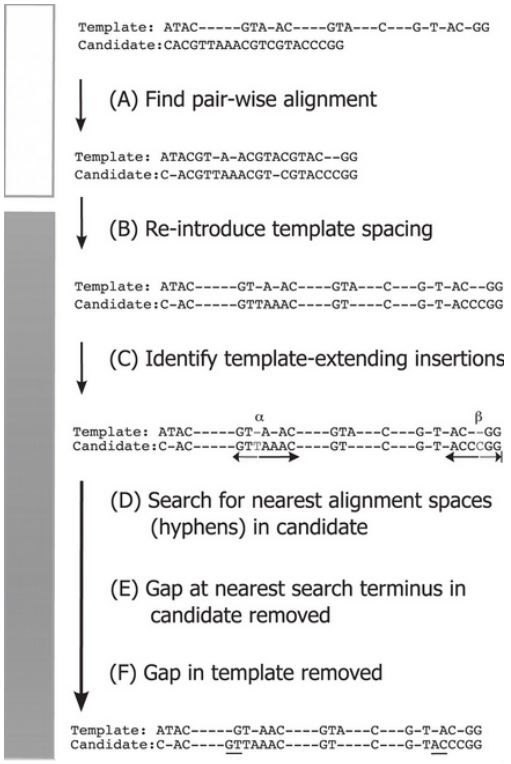


Figure 1.10: Example of NAST compression of a BLAST pairwise alignment using a 38 character aligned template. Figure from [27].

Others Other common alignment methods that we just mention are: MAFFT, which compute a multiple sequence alignment based on the fast Fourier transform (FFT); T-Coffee, which uses a progressive approach; INFERNAL, which tries to be more accurate and more able to detect remote homologous modeling sequences structure; mothur, through which one can do three different kinds of alignments: blastn (local), gotoh (global), and needleman (global).

1.3.2 Clustering methods

Besides alignment, another important step in sequence analysis is that of clustering sequences into OTUs (Operational Taxonomic Units). An OTU is a cluster of similar sequences, within a user defined threshold.
Clustering into OTU will be exploited in 16S rRNA sequencing analysis. In fact, for how these sequences are (see appendix C), a cluster of similar elements will correspond to bacteria in the same taxon at a particular taxonomic level.

BLAST BLAST first aligns the sequences using the homonymous method and then computes a single-linkage clustering [19], that is one of several methods of agglomerative hierarchical clustering. In the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined. The definition of 'shortest distance' is what differentiates between the different agglomerative clustering methods. In single-linkage clustering, the link between two clusters is made by a single element pair, namely those two elements (one in each cluster) that are closest to each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved. The method is also known as nearest neighbor clustering.
However, this method has different drawbacks. First of all, with the single-linkage clustering there will be the so-called chaining phenomenon, which refers to the gradual growth of a cluster as one element at a time gets added to it. This may lead to impractically heterogeneous clusters and difficulties in defining classes that could usefully subdivide the data. Moreover, BLAST is not really efficient in clustering divergent sequences, it can yield one-sequence clusters and it has a high dependency on the parameters choice (similarity threshold, identity percentage, alignment length). Finally, this algorithm is quite slower than other methods since it compares each sequence with all the others, a fact that makes it not suitable for large databases.

CD-HIT CD-HIT [1] has the main advantage of having ultra-fast speed. It can be hundreds of times faster than other clustering programs, like BLAST. Therefore it can handle very large databases. The main reason for this is that, unlike BLAST, which compute the all vs all similarities, CD-HIT can avoid many pairwise sequence alignments exploiting a short word filter.
CD-HIT uses greedy incremental clustering algorithm method. Briefly, sequences

are first sorted in order of decreasing length. The longest one becomes the representative of the first cluster. Then, each remaining sequence is compared to the representatives of existing clusters. If the similarity with any representative is above a given threshold, it is grouped into that cluster. Otherwise, a new cluster is defined with that sequence as the representative.
Here is how the short word filter works. Two strings with a certain sequence identity must have at least a specific number of identical words. For example, for two sequences to have 85% identity over a 100-residue window they have to have at least 70 identical 2-letter words, 55 identical 3-letter words, and 25 identical 5-letter words. By understanding the short word requirement, CD-HIT skips most pairwise alignments because it knows that the similarity of two sequences is below certain threshold by simple word counting.
A limitation of short word filter is that it can not be used below certain clustering thresholds, where the number of identical k-letter words could be zero (see fig.1.11).

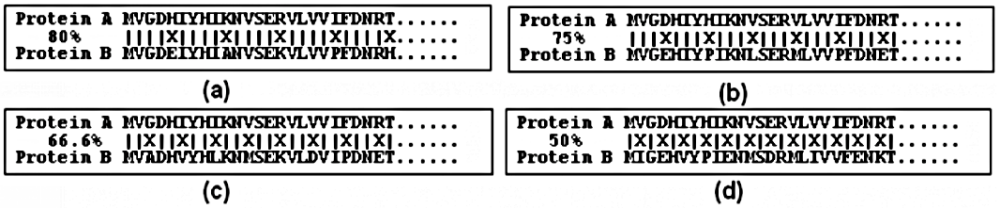


Figure 1.11: Short word filtering is limited to certain clustering thresholds. Evenly distributed mismatches are shown in alignments with 80%, 75%, 66.67% and 50% sequence identities. The number of common 5-letter words in (a), 4-letter words in (b), 3-letter words in (c), and 2-letter words in (d) can be zero. Figure from [1].

Another drawback of the algorithm is that it can happen that a sequence is more similar to a certain sequence but is put in the cluster of another one because it is compared first with this last one. Finally, two additional limitations are that CD-HIT does not give a hierarchical relation among clusters, and that it can yield one-sequence clusters.

Mothur There are three main steps in mothur [10]:

- aligns sequences;
- computes distances matrix;

- clustering.

For the clustering step mothur can use three different methods:

- nearest neighbor: each of the sequences within an OTU are at most $X\%$ distant from the most similar sequence in the OTU;

- furthest neighbor: all of the sequences within an OTU are at most $X\%$ distant from all of the other sequences within the OTU;

- average neighbor: this method is a middle ground between the other two algorithms.

Prefix/Suffix These methods [Qiime team, unpublished] collapse sequences which are identical in their first and/or last bases (i.e., their prefix and/or suffix). The prefix and suffix lengths are provided by the user and default to 50 each.

Trie Trie [Qiime team, unpublished] collapses identical sequences and sequences which are subsequences of other sequences.

USEARCH USEARCH [29] creates 'seeds' of sequences which generate clusters based on percent identity, filtering low abundance clusters. USEARCH can perform *de novo* or reference based clustering.

UCLUST UCLUST [2] is a method based on USEARCH. Its main advantages over previous methods are that it is faster, it uses less memory, it has an higher sensitivity and it is able to classify bigger datasets. We will describe this algorithm in more detail, since it is the one that we will use in our analysis.
The core step in the UCLUST algorithm is searching a database stored in memory. UCLUST performs *de novo* clustering by starting with an empty database in memory. Query sequences are processed in input order. If a match is found to a database sequence, then the query is assigned to its cluster (first figure below), otherwise the query becomes the seed of a new cluster (second figure below). Of course the first sequence in the input file will be the first seed of the database.

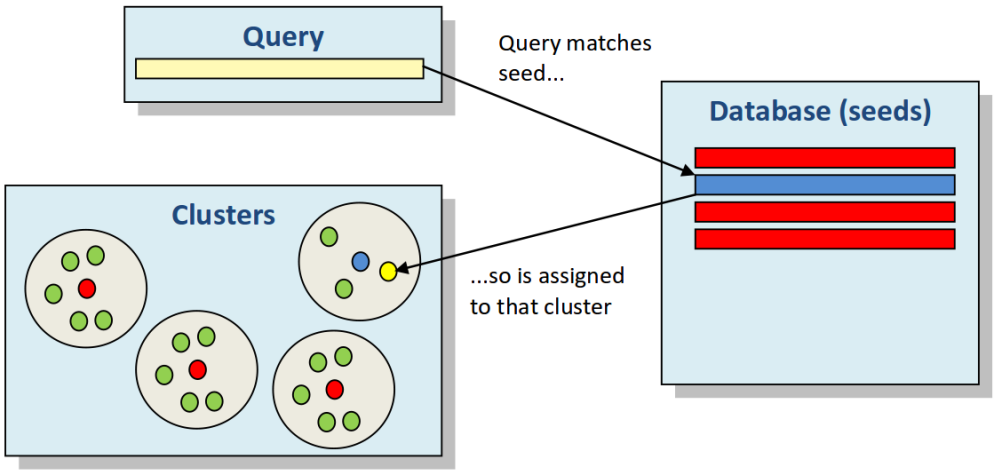


Figure 1.12: Schematic representation of the working of UCLUST if the query sequence matches a seed. Figure form [2].

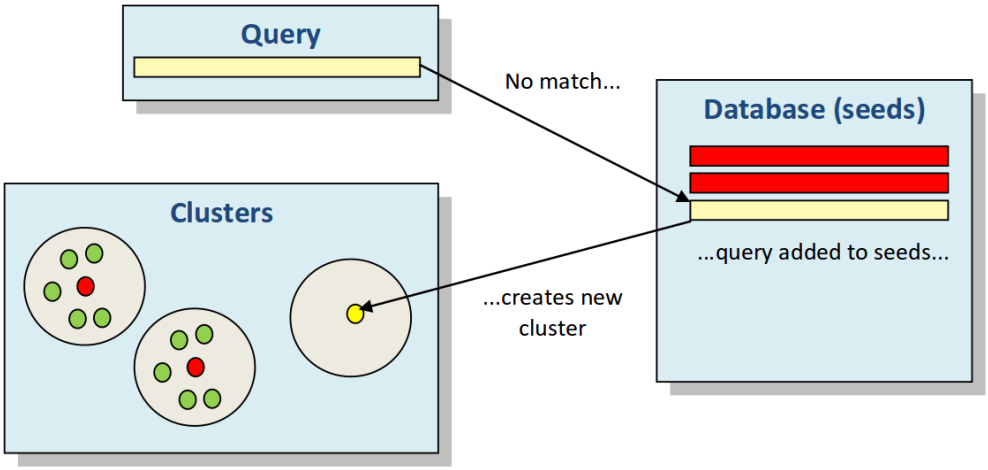


Figure 1.13: Schematic representation of the working of UCLUST if the query sequence does not match any seed. Figure form [2].

In this procedure, we say that a query sequence matches a database sequence if their similarity is high enough. Similarity is calculated from a global alignment, i.e. an alignment that includes all letters from both sequences. This differs from BLAST and most other database search programs, which search for local matches. The minimum identity is set in QIIME by the -s option, e.g. -s 0.97 means that

the global alignment must have at least 97% similarity. Similarity is computed as the number of matching (identical) letters divided by the length of the shortest sequence.
Let us observe that only seeds need to be stored in memory (because other cluster members do not affect how new query sequences are processed). This is an advantage for large datasets because the amount of memory needed and the number of sequences to search are reduced. However, this design may not be ideal in some scenarios because it allows non-seed sequences in the same cluster to fall below the identity threshold.
By default UCLUST stops searching when it finds a match. Usually UCLUST finds the best match first, but this is not guaranteed. If it is important to find the best possible match (i.e., the database sequence with highest similarity), then you can increase the --max_accepts option, even if in QIIME, the default value is 20, that is already increased compared to the default UCLUST value, that is 1.
UCLUST also stops searching if it fails to find a match. By default, it gives up after 8 failed attempts. Database sequences are tested in an order that correlates well (but not exactly) with decreasing similarity. This means that the more sequences get tested, the less likely it is that a match will be found later, so giving up early does not miss a potential hit very often. You can set the maximum number to try using the --max_rejects option, that in QIIME is 500 by default. With very high and very low similarity thresholds, increasing maxrejects can significantly improve sensitivity. Here, a rule of thumb is that low similarity is below 60% for amino acid sequences or 80% for nucleotides, high similarity is 98% or more.
By default, the target sequences are rejected if they have too few unique words in common with the query sequence. The threshold is estimated using heuristics. This improves speed, but may also reduce sensitivity. In QIIME there is the possibility to change the length of these words and to disable this option with the --word_length command.
An 'optimal' variant of the algorithm can be used specifying -A (--optimal_uclust), which is equivalent to setting --max_accepts and --max_rejects to 0 and to disable the rejection due to few words in common. This guarantees that every seed will be aligned to the query, and that every sequence will therefore be assigned to the highest-similarity seed that passes the similarity threshold (t). All pairs of seeds are guaranteed to have similarity $< t$. The number of seeds is guaranteed to be the minimum that can be discovered by greedy list removal, though it is possible that the number of clusters could be reduced by using a different set of seeds.
An 'exact' variant of the algorithm is selected by -E (--exact_uclust), which is equivalent to setting --max_accepts to 1, --max_rejects to 0 and to disable the rejection due to few words in common. This guarantees that a match will be found if one exists, but not that the best match will be found. The exact and optimal variants are guaranteed to find the minimum possible of clusters and both guarantee

that all pairs of seeds have identities $< t$. Exact clustering will be faster, but may have lower average similarity of non-seeds to seeds.
UCLUST supports a rich set of gap penalty options, even if in QIIME there is not a direct way to change them, probably because the default settings are considered optimized. By default, terminal gaps are penalized much less than interior gaps, which is typically appropriate when fragments are aligned to full-length sequences.
By default, UCLUST seeks nucleotide matches in the same orientation (i.e., plus strand only). You can enable both plus and minus strand matching by using -z (--enable_rev_strand_match). This command approximately doubles memory use but results in only small increases in execution time.
As we said at the beginning of this subsection, in UCLUST query sequences are processed in input order. This means that if a query was similar to more than a seed within the threshold, it will be put in the cluster of the sequence that was first in input order. Input sequences should therefore be ordered so that the most appropriate seed sequence for a cluster is likely to be found before other members. For example, ordering by decreasing length is desirable when both complete and fragmented sequences are present, in which case full-length sequences are generally preferred as seeds since a fragment may attract longer sequences that are dissimilar in terminal regions which do not align to the seed. In other cases, long sequences may make poor seeds. For example, with some high-throughput sequencing technologies longer reads tend to have higher error rates, and in such cases sorting by decreasing read quality score may give better results.
By default, UCLUST checks that input sequences are sorted by decreasing length, unless -D (--suppress_presort_by_abundance_uclust) is specified. This check can be disabled by specifying the -B (--user_sort) option, which specifies that input sequences have been pre-sorted in a way that might not be decreasing length.

1.3.3 Distances

mothur For computation of the distance matrix we will use mothur's command $dist.seqs$ [3]. This algorithm is well optimized, since the distances are not stored in RAM, but they are printed directly to a file. Furthermore, it is possible to ignore large distances that one might not be interested in.
To run $dist.seqs$ an alignment file must be provided in fasta format, so sequences should be aligned before computing their distances.
By default an internal gap is only penalized once, a string of gaps is counted as a single gap, terminal gaps are penalized (there is some discussion over whether to penalize them or not), all distances are calculated, and only one processor is used. You can change all these option through the corresponding commands.

The distances are computed as in the following example.
SequenceA: ATGCATGCATGC
SequenceB: ACGC - - - CATCC
Here, there would be two mismatches and one gap. The length of the shorter sequence is 10 nt, since the gap is considered as a single position. Therefore the distance would be 3/10 or 0.30.

1.3.4 Taxonomic assignment

RDP classifier For 16S rRNA taxonomic assignment, the most common algorithm is the RDP classifier [14].
The RDP Classifier is distributed with a pre-built database of assigned sequences, which is used by default. Each rRNA query sequence is assigned to a set of hierarchical taxa using a naive Bayesian rRNA classifier. The classifier is trained on the known type strain 16S sequences (and a small number of other sequences representing regions of bacterial diversity with few named organisms). The frequencies of all sixty-four thousand possible 8-base subsequences (words) are calculated for the training set sequences in each of the approximately 880 genera.
When a query sequence is submitted, the joint probability of observing all the words in the query can be calculated separately for each genus from the training set probability values. Using the naive Bayesian assumption, the query is most likely a member of the genera with the highest probability. In the actual analysis, the algorithm randomly selects only a subset of the words to include in the joint probability calculation, and the random selection and probability calculation is repeated for 100 trials. The number of times a genus is most likely out of the 100 bootstrap trials gives an estimate of the confidence in the assignment to that genus. For higher-order assignments, the algorithm sums the results for all genera under each taxon.
For each rank assignment, the Classifier automatically estimates the classification reliability using bootstrapping. Ranks where sequences could not be assigned with a bootstrap confidence estimate above the threshold are displayed under an artificial 'unclassified' taxon. The default threshold is 80%.
For partial sequences of length shorter than 250 bps (longer than 50 bps), a bootstrap cutoff of 50% was shown to be sufficient to accurately classify sequences at the genus level, and to provide genus level assignments for higher percentage of sequences (fig.1.14) [26].

Variable region	V3			V6			V4		
Bootstrap cutoff (≥)	0%	50%	80%	0%	50%	80%	0%	50%	80%
Fraction of sequences classified to genus	100%	92.4%	82.3%	100%	73.5%	40.4%	100%	97.0%	87.9%
Fraction of sequences correctly classified to genus	92.0%	95.0%	98.1%	79.0%	96.5%	98.7%	92.8%	94.5%	95.7%

Figure 1.14: Of 7208 full-length 16S reference sequences from the human gut 6054 were classified at genus-level with 80% bootstrap support. With these full-length assignments as references the V3, V4 and V6 regions were extracted and re-classified at three different bootstrap thresholds, and compared with the full-length classification (last row). Figure form [26].

We can choose to use 50% as bootstrap cut-off since the accuracy is closest to the one with 80% cut-off, and the total number of sequences that could be assigned to genus level is closest to that obtained without any cut-off threshold imposed.

Others Other methods exploited in QIIME are the following [13].

- **BLAST**. Taxonomy assignments are made by searching input sequences against a BLAST database of pre-assigned reference sequences. If a satisfactory match is found, the reference assignment is given to the input sequence. This method does not take the hierarchical structure of the taxonomy into account, but it is very fast and flexible.

- **RTAX**. Taxonomy assignments are made by searching input sequences against a fasta database of pre-assigned reference sequences. All matches are collected which match the query within 0.5% identity of the best match. A taxonomy assignment is made to the lowest rank at which more than half of these hits agree.

- **mothur**. The mothur software provides a naive bayes classifier similar to the RDP Classifier. A set of training sequences and id-to-taxonomy assignments must be provided. Unlike the RDP Classifier, sequences in the training set may be assigned at any level of the taxonomy.

In their study, Claesson et al. [26] compared different algorithms for taxonomic assignment and found out that the Greengenes and RDP-classifier produced the most accurate and stable results, especially for gut communities. Furthermore, the

RDP-classifier resulted more than 30 times faster than the Greengenes classifier. Thus, they chose the RDP classifier due to its documented accuracy and stability, straightforward usage, independence of sequence alignments, high speed, and suitability for very large datasets generated by next-generation sequencing technologies, and so we will do.

Chapter 2

Gut microbiota and microbioma

In this chapter we will provide an insight of the gut microbiota as a biomedical issue, in order to give an idea of the importance of this ecosystem and its biodiversity for our health, reminding how the latest sequencing techniques and ecological theories are necessary for this purpose.

2.1 Metagenomics

Classical microbiology relied largely on the culturing and analysis of microbes isolated from environmental samples. However, in the 1990s studies based on the real cell counts using microscopy techniques and 16S rRNA phylogenetic profiling, estimated that the currently cultivatable microorganisms represent only a small fraction (less than 1%) of the total microbes within a given habitat [38]. Thus traditional clonal culture techniques result biased and cannot access the vast majority of organisms within a community.
Metagenomics is a mean to overcome these issues by capturing and analyzing the genetic material of the entire microbial community (i.e. the metagenome), relying on the cultivation-independent extraction of total environmental DNA.
Thus metagenomics exploits sequencing techniques (see chapter 1) to answer questions such as: how many different species inhabit a particular environment, what is the genomic potential of that community (i.e. which genes, functions or pathways are present), which species are responsible for which activities, and how does the community change over time and under different environmental conditions [38].
In particular, in our work we are going to analyze gut microbiota data of next-generation sequencing and to model them through ecological theories to give biodiversity informations (see chapter 4).

2.2 Human microbiota

As reported in [38], the human body is home to roughly 10 times more microbial cells than human cells. These commensal and not pathogenic microorganisms (called human microbiota) come from all three domains of life: bacteria, archaea, and eukaria, as well as viruses, and are found mostly in the gastrointestinal tract but also along the skin surface, oral and nasal cavities, and urogenital tracts. The collective genomes of all these symbiotic microorganisms (called human microbiome) constantly interacts with the human genomes, making humans 'superorganisms' harboring these two integrated genomes. It is through their interaction with our living environment that the human health phenotype is defined; than it is this interaction that we should consider in the study of systemic diseases.
Through the diet we influence the composition of bacteria living environment and therefore of bacteria population in our gut, yielding to a change in the metabolites production by the microbiota, that can get into our bloodstream via a normal route enterohepatic circulation or through partially impaired gut barrier and eventually influence human health. In particular changing patterns of food consumption has been closely linked with the dramatic increase in the incidence of obesity, diabetes, and cardiovascular diseases, linked with variations in gut microbiota distribution. Furthermore gut microbiota exhibits significant changes in response to health changes, even in the early phase in which these are not yet detectable, like during the development of precancerous lesions in the gut. These features make gut microbiota both a biomarker for health changes and a target for nutritional/medicinal interventions in chronic diseases.

2.3 Gut microbiota and metabolic diseases

Gut microbiota - normal functioning

The human gut is composed of four main regions: the oesophagus, the stomach, the small intestine, and the large intestine, constituted by the caecum and the colon. Through molecular analysis of gut microorganisms sampled through biopsies or luminal content analysis, researchers have obtained an outline of gut microbial diversity.
The first results from these analyses indicated that the same bacterial phyla tend to predominate in the stomach, small intestine, caecum and large intestine. Thus

more than 98% of all species detected belong to four phyla, and on average Firmicutes (64%) and Bacteroidetes (23%) lead the way in terms of abundance in front of the Proteobacteria (8%) and the Actinobacteria (3%).
Members of Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, Fusobacteria and TM7 (an uncultured bacterial phylum) were predominantly detected in the upper digestive tract (oesophagus), and 41 genera were detected within these six phyla.
The human stomach microbiota is dominated by Firmicutes and Bacteroidetes and contains at low relative abundance representatives of the phyla Actinobacteria, Fusobacteria, TM7, Deferribacteres, and Deinococcus/Thermus.
In the distal gut (caecum, colon, and fecal samples) the predominant bacterial phyla are the Firmicutes and Bacteroidetes, in addition to one Archaeal species, Methanobrevibacter smithii. The remaining phyla such as Proteobacteria, Actinobacteria, Fusobacteria, TM7 and Verrucomicrobia, are present at much lower frequencies. Despite the restricted number of dominant bacterial phyla found in the distal gut, each individual harbors a remarkable number of species. A recent study on the fecal microbiota of an adult monozygotic female twin pair revealed an estimated 800-900 bacterial species in each co-twin, less than half of which were shared by both individuals. Expanding this analysis to include a shallower sampling of 21 fecal samples obtained from 54 mono- and dizygotic twin pairs and their mothers revealed > 4000 species-level bacterial phylogenetic types (phylotypes). However, of the 134 bacterial species whose relative abundance was $> 0.1\%$ in at least one fecal community, only 37 were detected in more than half of the analyzed samples [38]. These studies can make us understand how variable and subjective is our microbioma.

Gut microbiota - functioning in metabolic diseases

As reported previously, gut microbiota is influenced by and in turn influence the health state of the host human organism. Recent works have shown in particular the relation of gut microbiota with metabolic diseases, and in particular how the disruption of gut microbiota by high fat-diet may play a pivotal role in the onset and progression of obesity and insulin resistance, that constitute the early stage of these kind of diseases.
Gnotobiotic (i.e. with only certain known strains of bacteria and other microorganisms present) mice model have been highly instrumental in the elucidation of the mechanisms for gut microbiota getting involved in obesity development. It was shown that in contrast to mice with a gut microbiota, germ-free animals are protected against the obesity that develops after consuming a Western-style, high-fat, sugar-rich diet [23].

Furthermore, in [44], the authors showed that the transfer of gut microbiota from genetically obese mice or from high-fat diet-induced obese mice to germ-free wild-type lean mice lead to significant accumulation of fat in the latter, indicating that gut microbiota are not only necessary but also sufficient to induce obesity in the host animals.
Gut microbiota participates in obesity development in two ways [38]: (1) ferments plant polysaccharides into short chain fatty acids, thus helping the host extract more calories from otherwise indigestible food components; (2) distorts host energy metabolism by directly regulating relevant genes, for example by suppressing the expression of the fasting-induced adipogenic factor (Fiaf) gene, that is required for fatty acid oxidation, consequently limiting it.
Gut microbiota can also stimulate genes involved in triglycerides synthesis in the liver (associated with the development of insulin resistance) thus acting on both sides of the energy equation and transforming the host animals into a highly efficient fat-making and storage machine [23].
Another distinct but complementary mechanism for gut microbiota getting involved in development of obesity and insulin resistance is by way of provoking a low-grade, systemic and chronic inflammatory condition, a key underlying pathological condition in the development of these metabolic diseases.
An important factor in this process is an endotoxin called lipopolysaccharide (LPS). LPS is a major component of the outer membrane in Gram-negative bacteria [1]. Being an endotoxin, if LPS overcomes the gut mucosal barrier and enters the circulatory system, it causes a toxic reaction, stimulating an immune response (it activates B cells and induces macrophage and other cells to release interleukin-I and interleukin-6, tumor necrosis factor, and other factors), with the sufferer developing a high temperature, high respiration rate, and low blood pressure.
As reported in [25], recent works showed that LPS was increased 2-3 times in the blood of high-fat diet fed animals, which showed low-grade systemic chronic inflammation comparable to what has been found in human subjects. LPS resulted responsible for the onset of metabolic diseases, since a continuous subcutaneous low-rate infusion of LPS induced most, if not all, of the features of metabolic diseases and since the corresponding LPS receptor CD14 knockout mouse resisted the occurrence of the diseases.
Furthermore a subcutaneous injection of comparable amount of LPS into the bloodstream of mice fed on normal chow diet made the otherwise lean and healthy

[1]Bacteria can be classified, based on their cell wall structure, through the Gram stain test in which a counterstain (commonly safranin) is added after the crystal violet. Gram-positive bacteria will retain the crystal violet dye when washed in a decolorizing solution, unlike Gram-negative bacteria which will not absorb the gram stain thanks to the thick lipid bilayer membrane (whose outer layer contains LPS). Compared with Gram-positive bacteria, Gram-negative bacteria are more resistant against antibodies, because of their impenetrable wall.

animals start to develop inflammation and eventually became obese and insulin resistant.
LPS levels resulted closely correlated with the Gram negative-to-Gram positive ratio within the gut. In [25] and [24], the authors pointed out how high-fat diet dramatically changed the gut microbiota content, leading to endotoxemia, and how dietary fibers, which reduce the impact of high-fat diet on the occurrence of the metabolic diseases, normalized the Gram negative-to Gram positive ratio and consequently the plasma endotoxemia. In particular high-fat feeding decreased the number of Bifidobacteria, a group of bacteria that has been shown to reduce intestinal LPS levels and to improve the mucosal barrier protecting function.
About this, the authors found also out that high-fat diet-induced metabolic endotoxemia depended on a mechanism involved in the control of gut permeability by gut bacteria (in fact antibiotic treatment restored normal plasma LPS values). In particular they proved that high-fat feeding dramatically increases intestinal permeability by a mechanism associated with a reduced expression of epithelial tight junction proteins such as ZO-1 and Occludin. The importance of this finding is due to the fact that the enhancement in intestinal permeability led to metabolic endotoxemia, that positively correlated with inflammation, oxidative stress and macrophage infiltration.
In [52], Zhang et al., used DNA fingerprinting and bar-coded pyrosequencing of 16S rRNA genes to profile gut microbiota structures and identified sulphate-reducing bacteria in family Desulfovibrionaceae as the potentially important endotoxin producers, whose abundance changes were associated with the development of metabolic syndrome in mice. Member of this family are Gram-negative, opportunistic pathogen, endotoxin producers and are also capable of reducing sulphate to H_2S, damaging the gut barrier.
Furthermore, in [52] the authors showed that 56% of structural variations of gut microbiota can be attributed to diet types while only 12% to host genetic mutation. Animals with no genetic defect can develop sever obesity and insulin resistance by taking excess amount of high-fat diet. That means, animals do not need to harbour a genetic defect to develop metabolic diseases. The disruption of gut microbiota by a high-fat diet would be sufficient to distort the host energy metabolism and provoke inflammation, causing severe adiposity and tissue damage.
From all these experiments we can finally deduce that a high-fat diet disruption of gut microbiota plays an essential mediator role in the inflammatory condition responsible for onset and progression of obesity, as well as of related metabolic diseases such as type 2 diabetes and cardiovascular diseases.

Gut microbiota - functioning in type 2 diabetes

In [36], the authors, using pyrosequencing of the V4 region of the 16S rRNA gene and qPCR, demonstrated that type 2 diabetes is associated with compositional changes in the intestinal microbiota mostly apparent at phylum and class level.
In agreement with other results obtained for overweight persons, they found out a significant lowering in the relative abundance of Firmicutes in diabetics, while the proportion of Bacteroidetes and Proteobacteria was somewhat higher.
Bacterial group that distinguished the diabetic from the non-diabetic microbioma included Bacteroides-Prevotella group versus class Clostridia and C. coccoides-E.rectale group, which ratios were significantly higher in diabetic persons. These results are supported by previous studies showing reduction in Bacteroides-Prevotella spp. related to a strong decrease of metabolic endotoxemia and inflammation in type 2 diabetes mice. Accordingly, a significant reduction in Clostridium ssp, C. coccoides and an increase in the Bacteroides-Prevotella group along with body weight loss have been observed in human studies.
Larsen et al. also reported significantly higher levels of Bacilli and the Lactobacillus group in diabetic subjects compared to controls, an important finding since Genus Lactobacillus represents a heterogeneous group with well documented immunomodulating properties and might potentially contribute to chronic inflammation in diabetic subjects.
As reported previously, in an obesity study, using mice models, Cani et al. found a connection between metabolic diseases and the presence of Gram-negative bacteria in the gut. Accordingly, also the intestinal microbiota across type 2 diabetes subjects resulted relatively enriched with Gram-negative bacteria, belonging to the phyla Bacteroidetes and Proteobacteria.
To carry out further analysis on gut microbial content in patients with type 2 diabetes, Qin et al. developed a protocol for a metagenome-wide association study (MGWAS) and undertook a two-stage MGWAS based on deep shotgun sequencing of the gut microbial DNA from 345 Chinese individuals.
In their work emerged that T2D patients had only a moderate degree gut bacterial dysbiosis, while functional annotation analyses indicated a decline in butyrate-producing bacteria, which may have a protective role against several types of deaseas and be metabolically beneficial, and an increase in several opportunistic pathogens. The authors revealed also an enrichment of other microbial functions conferring sulphate reduction and oxidative stress resistance and finally suggested that there is a ‘functional dysbiosis’ rather than a specific microbial species that has a direct association with T2D (see fig. 2.1), and this underlines the need of understanding the dynamics of the whole bacteria population which constitutes the gut microbiota.

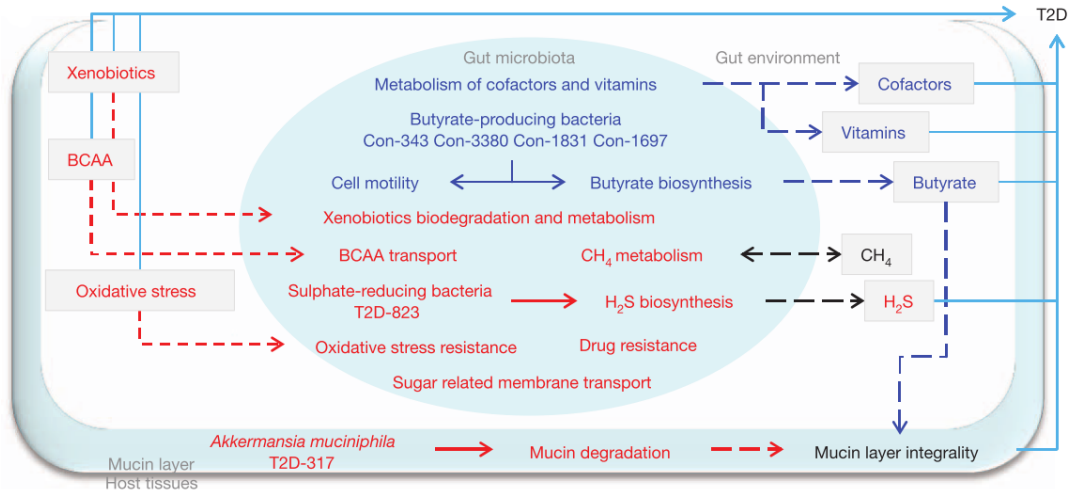


Figure 2.1: A schematic diagram showing the main functions of the gut microbes that had a predicted T2D association. Red text denotes enriched functions in T2D patients; blue text denotes depleted functions in T2D patients; black text denotes an uncertain functional role relative to T2D. Figure from [40].

In this chapter we showed how the gut microbiota dysbiosis and functioning alteration is of crucial importance in many metabolic diseases, from obesity to type 2 diabetes. Thus, you should have understood the reason for which we want to obtain a dynamical model to describe the gut microbiota population biodiversity and evolution. In the next two chapters we will face the dynamical model proposed by Volkov et al. [49] to explain the biodiversity of other ecological systems, introducing also the required mathematical tools. Finally we will show our application of these dynamical models to next-generation sequencing data of gut microbiota.

Chapter 3

The Chemical Master Equation

In this chapter we are going to give an overview on the Chemical Master Equation, that we will then exploit in chapter 4 to describe ecological communities.

The Master equation is an equation that describes the time-evolution of the probability of a system to be in a specific configuration, driven by a memoryless process of transition between states (markovian process). The systems considered are those that can be modeled as being in exactly one of countable number of states at any given time, and whose switching between states is treated probabilistically. This approach is particularly useful to describe biological phenomenon, in which a deterministic approach would be incorrect, for example when the number of molecules in the system is small and thus the fluctuations are not negligible. In these cases a mean field approach is not correct and one needs to introduce some kind of noise and to treat the system stochastically.
Thus, the Chemical Master Equation is used to describe different biological situations, from the enzymatic reactions inside the cell (which involve in fact a small number of molecules), to the protein production, to the bet hedging strategies of some bacteria that, in order to guarantee the survival of the species to abrupt climate changes, generate a variable offspring that could survive in different environments.

3.1 Markov processes

Of course, the main concept behind the master equation is that of markovian process. So, let us now define what a markovian process is, starting from the definitions of stochastic variable and stochastic process, referring to [46].
A stochastic process is defined as a function of the time t and a stochastic variable

X where for each value of X we observe a different realization of the stochastic process.

Stochastic variables A stochastic variable is an object X defined by

- a set of possible values (called range, set of states, sample space, or phase space) that may be discrete or continuous and mono- or multi-dimensional;
- a probability distribution function $P(x)$ over this set, that should of course satisfy the properties $P(x) \geq 0$ and $\int P(x)dx = 1$.

Stochastic processes Once a stochastic variable X has been defined, an infinity of other stochastic variables derives from it, namely all quantities Y that are defined as functions of X by some mapping f. These quantities Y may be any kind of mathematical object, in particular also functions of an additional variable t,

$$Y_X(t) = f(X, t). \tag{3.1}$$

Such a quantity $Y(t)$ is called random function, or, if t stands for time, a stochastic process. On inserting for X one of its possible values x, we obtain a so called sample function, or realization of the process $Y_x(t) = f(x, t)$. We refer to a stochastic process as an ensemble of these sample functions. The process Y can describe any kind of phenomenon, like the state of a subatomic particle moving through matter, the position of a Brownian particle, the number of molecules of each kind in a chemical reaction, or the number of individuals of a certain species in an ecological system.
We can describe the probability of observing a specific value of the function Y at a given time t as the measure of the ensemble of values of X for which the function Y gives the value y at time t

$$P(y, t) = \int_{-\infty}^{\infty} \delta(Y(X, t) - y)dX. \tag{3.2}$$

Markov processes A Markov process is defined as a stochastic process Y in which there is no relationship between the value of Y at a certain time and its value at the previous moments, so that the probability of observing y_n at the time t_n, given the observation of the values of Y at the previous times $t_1, \cdots, t_{n-1}$, only depends from the value of Y at t_{n-1}

$$P_{1|n-1}(y_n, t_n|y_1, t_1; \cdots; y_{n-1}, t_{n-1}) = P_{1|1}(y_n, t_n|y_{n-1}, t_{n-1}), \tag{3.3}$$

where the conditional probability density $P_{1|1}$ is the so called transition probability.
Therefore in a markovian process we have a very simple relationship: the probability distribution of observing the value y_n at time t_n, is a function of only the state of the system at the time t_{n-1} and is not affected by any knowledge of the values at earlier times. In other words, markovian systems are memory-less, meaning that they lose any kind of information of their state before the present value.
A Markov process is fully determined by the two functions $P_1(y_1, t_1)$ and $P_{1|1}(y_2, t_2|y_1, t_1)$, in fact we can find successively all P_n from the iteration of

$$\begin{aligned} P_3(y_1, t_1; y_2, t_2; y_3, t_3) &= P_2(y_1, t_1; y_2, t_2) P_{1|2}(y_3, t_3|y_1, t_1; y_2, t_2) \\ &= P_1(y_1, t_1) P_{1|1}(y_2, t_2|y_1, t_1) P_{1|1}(y_3, t_3|y_2, t_2), \end{aligned} \tag{3.4}$$

and this is the property that makes Markov processes manageable and consequently so useful in applications.

The Chapman-Kolmogorov equation To understand the meaning of the Chapman - Kolmogorov equation we can start from the following example about the discrete case. Let us consider a rat in a maze with four cells, indexed 1 - 4, plus the outside (freedom), indexed by 0 (that can only be reached via cell 4) as showed in fig.3.1.

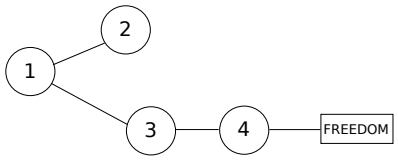


Figure 3.1: Rat maze's diagram.

The rat starts initially in a given cell and then takes a move to another cell, continuing to do so until finally reaching freedom. We assume that at each move (transition) the rat, independently from the past, is equally likely to chose from among the neighboring cells (so we are assuming that the rat does not learn from past mistakes). This then yields a Markov chain, where y_n denotes the cell visited right after the n-th move. In our case, for example, whenever the rat is in cell 1, it moves next (regardless of its past) into cell 2 or 3 with probability $1/2$: $P_{1,2} = P_{1,3} = 1/2$. Then, the probability that the rat, starting initially in cell 1, is back in cell 1 two steps later, denoted as $P_{11}^2 = P(y_2 = 1|y_0 = 1)$, is given by the

probability for the rat to go to cell 2 and then back to cell 1, plus the probability to go to cell 3 and then back to cell 1, that is $P_{11}^2 = P(y_1 = 2, y_2 = 1|y_0 = 1) + P(y_1 = 3, y_2 = 1|y_0 = 1) = 1/4 + 1/4 = 1/2$.
We can demonstrate that in general it holds that $P^{(n)} = P^n = P \times P \times \cdots \times P$, $n \geq 1$, and that the probability of a transition from the state y_n (time t_n) to the state y_{n+m} (time t_{n+m}), $P_{y_n,y_{n+m}}$, follows the Chapman-Kolmogorov equation

$$\begin{aligned} P_{y_n,y_{n+m}}^{n+m} &= P(y_{n+m}, t_{n+m}|y_n, t_n) \\ &= \sum_{y_k \in S} P_k(y_k, t_k|y_n, t_n) P_m(y_m, t_m|y_k, t_k), \end{aligned} \tag{3.5}$$

$\forall n \geq 0, \forall m \geq 0$, with $y_n, y_m \in S$, where S is the state space. The above equation is derived by first considering in what state the chain (i.e. the process) is at time n. Given as initial state the value y_n, $P_k(y_k, t_k|y_n, t_n)$ is the probability that the state at time t_k is y_k. But then, given y_k, the future after time t_n is independent of the past, so the probability that the chain m time units later (at time t_{n+m}) will be in state y_m is $P(y_{n+m}, t_{n+m}|y_k, t_k)$, and thus, from the independence of the probabilities, we have $P(y_k, t_k; y_{n+m}, t_{n+m}|y_0 = y_n, t_0 = t_n) = P_k(y_k, t_k|y_n, t_n) P_m(y_m, t_m|y_k, t_k)$. Summing up over all k yields the result.
A rigor proof of the Chapman-Kolmogorov equation is given by the expression below.

$$\begin{aligned} P_{i,j}^{n+m} &= P(X_{n+m} = j|X_0 = i) \\ &\overset{1}{=} \sum_{k \in S} P(X_{n+m} = j, X_n = k|X_0 = i) \\ &\overset{2}{=} \sum_{k \in S} \frac{P(X_{n+m} = j, X_n = k, X_0 = i)}{P(X_0 = i)} \\ &= \sum_{k \in S} \frac{P(X_{n+m} = j|X_n = k, X_0 = i) P(X_n = k, X_0 = i)}{P(X_0 = i)} \\ &\overset{3}{=} \sum_{k \in S} \frac{P(X_n = k, X_0 = i) P_{k,j}^m}{P(X_0 = i)} \\ &= \sum_{k \in S} P_{i,k}^n P_{k,j}^m, \end{aligned} \tag{3.6}$$

[1] Summing over all the elements in S we consider all the possible ways to go from i to j.

[2] From the definition of conditional probability $P(A|B) = \frac{P(A \bigcap B)}{P(B)}$.

[3] We used the Markov property to conclude that $P(X_{n+m} = j|X_n = k, X_0 = i) = P(X_n + m = j|X_n = k) = P(X_m = j|X_0 = k) = P_{k,j}^m$.

We can extend the Chapman-Kolmogorov equation to the continuous case integrating the identity 3.4 over y_2. For $t_1 < t_2 < t_3$ we obtain

$$P_2(y_1, t_1; y_3, t_3) = P_1(y_1, t_1) \int P_{1|1}(y_2, t_2|y_1, t_1) P_{1|1}(y_3, t_3|y_2, t_2) dy_2, \quad (3.7)$$

and dividing both sides by $P_1(y_1, t_1)$, we obtain the continuous form of the Chapman-Kolmogorov equation

$$P_{1|1}(y_3, t_3|y_1, t_1) = \int P_{1|1}(y_3, t_3|y_2, t_2) P_{1|1}(y_2, t_2|y_1, t_1) dy_2. \quad (3.8)$$

And this is an identity which must be obeyed by the transition probability of any Markov process.
As observed previously, a Markov process is fully determined by P_1 and $P_{1|1}$ because the whole hierarchy P_n can be constructed from them, but these two functions cannot be chosen completely arbitrarily, since the must obey two identities:

- the Chapman-Kolmogorov equation 3.8;
- the obvious relation $P_1(y_2, t_2) = \int P_{1|1}(y_2, t_2|y_1, t_1) P_1(y_1, t_1) dy_1$

Vice versa, any two nonnegative functions P_1 and $P_{1|1}$ that obey these conditions uniquely define a Markov process.

Stationary Markov processes Let us consider a closed and isolated physical system that has a quantity, or a set of quantities, $Y(t)$ which may be treated as a Markov process. If the system is in equilibrium (for some quantities, it is sufficient to have the system at the steady state), we can assert that $Y(t)$ is a stationary process. Stationary Markov processes are particularly interesting for describing equilibrium fluctuations.
For these processes, P_1 results independent of time and becomes the familiar equilibrium distribution of Y, and the transition probability $P_{1|1}$ does not depend on time's instants but only on the time interval. Under these conditions we can introduce the special notation

$$P_{1|1}(Y_2, t_2|y_1, t_1) = T_\tau(y_2|y_1), \quad (3.9)$$

with $\tau = t_2 - t_1$.
For $\tau, \tau' > 0$ we can then rewrite the Chapman-Kolmogorov equation as

$$T_{\tau+\tau'}(y_3|y_1) = \int T_{\tau'}(y_3|y_2) T_\tau(y_2|y_1) dy_2. \quad (3.10)$$

3.2 The Master Equation

The Master equation is an equivalent form of the Chapman-Kolmogorov equation for Markov processes, but it is easier to handle and more directly related to physical processes. It is a differential equation obtained by letting τ' to zero.
To derive the master equation let us consider a Markov process, which for convenience we take to be homogeneous, so that we can write T_τ for the transition probability, and let us see how $T_{\tau'}$ behaves as τ' tends to zero, calculating the first order Taylor expansion of $T_{\tau'}(y_3|y_2)$ for τ' small

$$\begin{aligned} T_{\tau'}(y_3|y_2) &= T(\tau'=0) + \partial_{\tau'}T_{\tau'}(\tau'=0)\cdot\tau' + O(\tau'^2) \\ &= (1-\alpha_0\tau')\delta(y_2-y_1) + \tau' W(y_2|y_1) + O(\tau'^2), \end{aligned} \tag{3.11}$$

where $W(y_3|y_2)$ is the transition probability per unit time from y_2 to y_3, also called transition rate, and where $T(\tau'=0) = (1-\alpha_0\tau')\delta(y_2-y_1)$ represents the fact that, for small time intervals, the system will not move too much away from the state y_2 and that the probability of remaining in that state will be slightly less than 1. This last statement comes from the term $(1-\alpha_0\tau')$, where α_0 is the normalization constant $\alpha_0(y_2) = \int W(y_3|y_2)dy_3$, that we obtain from

$$\begin{aligned} \int T_{\tau'}(y_3|y_2)dy_3 = 1 = 1 - \alpha_0(y_2)\tau' + \tau'\int W(y_3|y_2)dy_3 \\ \Rightarrow \alpha_0(y_2) = \int W(y_3|y_2)dy_3. \end{aligned} \tag{3.12}$$

Inserting the Taylor expansion 3.11 in the Chapman-Kolmogorov equation 3.8, we obtain for τ' small,

$$\begin{aligned} T_{\tau+\tau'}(y_3|y_1) &= \int \left(\delta(y_3-y_2)(1-\alpha_0(y_2)\tau')T_\tau(y_2|y_1) + \tau' W(y_3|y_2)T_\tau(y_2|y_1)\right) dy_2 \\ &= (1-\alpha_0(y_3)\tau')T_\tau(y_3|y_1) + \tau'\int W(y_3|y_2)T_\tau(y_2|y_1)dy_2, \end{aligned}$$

where we have applied the definition of δ.
Considering that $\frac{T_{\tau+\tau'}-T_\tau}{\tau'} \underset{\tau'\to 0}{\longrightarrow} \frac{d}{d\tau}T_\tau$, we have

$$\frac{d}{d\tau}T_\tau(y_3|y_1) = -\alpha_0(y_3)T_\tau(y_3|y_1) + \int W(y_3|y_2)T_\tau(y_2|y_1)dy_2. \tag{3.13}$$

Inserting $\alpha_0(y_3) = \int W(y_2|y_3)dy_2$, we obtain

$$\frac{d}{d\tau}T_\tau(y_3|y_1) = \int \left(W(y_3|y_2)T_\tau(y_2|y_1) - W(y_2|y_3)T_\tau(y_3|y_1)\right) dy_2, \tag{3.14}$$

where the first positive term of the integral represents the rate of jumping towards y_3 (influx), while the second negative term stands for the rate of jumping out from y_3 (efflux).
The equation 3.14 is called master equation and describes the variation of the probability of the system to be in a particular state, due to incoming and outgoing fluxes.
We may observe that for many systems it is much easier to determine the transition rate $W(y_2|y_1)$, through measurements or modeling, than the whole transition probability $T_\tau(y_2|y_1)$, that can instead be determined through the master equation.

3.3 Chemical Master Equation (CME)

The master equation 3.14 refers to a specific y_1 and t_1 and we can rewrite it removing all the redundant indexes, through the change of variables

$$y_3 \to y$$
$$y_2 \to y'$$

In this way we obtain the probability of observing the state y at the time t

$$\partial_t P(y,t) = \int [W(y|y')P(y',t) - W(y'|y)P(y,t)]\, dy'. \tag{3.15}$$

This equation represents an influx of probability to the state y from all the connected states y' and an efflux from y to every state y' to which it can move.
If the system state space is discrete, as when we work with a system with a discrete number of individuals or molecules, we can write the probability as $P_n(t)$ to represent the discreteness of the state space. In this case the master equation can be called Chemical Master Equation (referring to a chemical environment) and we can rewrite it, replacing integrals with sums, as

$$\partial_t P_n(t) = \sum_{n=0}^{\infty} [\lambda_{n',n} P_{n'}(t) - \lambda_{n,n'} P_n(t)], \tag{3.16}$$

where the λs are the discrete versions of the Ws of the continuous equation, with the origin and destination index exchanged, so that $\lambda_{n,n'}$ represents the probability flux from the state n to the state n'.
For a linear dynamic system (i.e. a system for which the effects superposition principle holds) we can further simplify equation 3.16, writing

$$\partial_t \overrightarrow{P}(t) = \Lambda \overrightarrow{P}(t), \tag{3.17}$$

where the matrix Λ is called the transition matrix and is defined as

$$\Lambda_{i,j} = \begin{cases} \Lambda_{i,j} = \lambda_{i,j}, \forall i \neq j \\ \Lambda_{i,i} = -\sum_{j \neq i} \lambda_{i,j} \forall i = j \end{cases} \tag{3.18}$$

where the second expression means that the probability of remaining in the state i is equal to minus the probability of exiting from i.

Properties of the transition matrix We can observe that Λ is a zero determinant matrix by construction, since $\sum_j \Lambda_{i,j} = 0 \; \forall i$. The zero determinant matrix represents the conservation of probability; in fact a determinant different from zero means that a certain amount of probability would be generated or destroyed, and this is absurd, since we are describing the system as a whole (i.e. closed). A zero determinant also means that there is at least one zero eigenvalue, since we can write the determinant of a matrix as the product of its eigenvalues. Furthermore the eigenvector corresponding to this null eigenvalue, is the so called stationary distribution, that is the distribution, given by $\partial_t P_n(t) = 0$, to which the stochastic process always converges, as long as the transition propensities λ are not function of time. If the system is fully connected (thus it cannot be broken into two non communicating pieces), the stationary distribution is guaranteed to be unique. Furthermore the stationary distribution will be positive, that is all its terms will be positive, and the sum of its components will be 1, being a probability distribution. To see these properties, let us consider the differential equation that describes our system:

$$\partial_t \overrightarrow{P}(t) = \Lambda \overrightarrow{P}(t). \tag{3.19}$$

The solution of this differential equation is, integrating by separation of variables, of the kind $\overrightarrow{P}(t) = \overrightarrow{P}(0) e^{\Lambda t}$. Introducing the eigenvectors u_α and the eigenvalues λ_α of the matrix Λ, defined by the equation $\Lambda u_\alpha = \lambda_\alpha u_\alpha$, we can rewrite the solution of the master equation as

$$P(t) = \sum_\alpha c_\alpha e^{\lambda_\alpha t} u_\alpha, \tag{3.20}$$

where the coefficients c_α are determined by the initial conditions.
From this expression we can firstly observe that if, for a certain state α, the eigenvalue is $\lambda_\alpha = 0$, then the P_α corresponding to this state is the stationary distribution.
We can also notice that the eigenvectors represent the direction of the decay of the probability, and the absolute value of the eigenvalues gives the velocity of this decay: a small $|\lambda_\alpha|$ stands for a slow decay, while a large $|\lambda_\alpha|$ corresponds to a

fast one. Furthermore, all the others eigenvalues will have a negative real part. In fact the probabilities $P_\alpha = u_\alpha e^{\lambda_\alpha t} c_\alpha$ are nonnegative numbers less than or equal to 1, and so positive values of λ_α are physically impossible since they would lead to exponentially growing probabilities, i.e. > 1. Under this point of view we can interpret the existence of the null eigenvalue observing that, if all the λ_αs were negative, than all the probabilities would go to zero for $t \to \infty$, while we know that the system will certainly be in one of the N states and that consequently the sum of the probabilities have to be 1 and not 0. In the limit $t \to \infty$ the probabilities of the states will tend to their equilibrium values $P_\alpha(\infty) = P_\alpha^{(eq)}$, thus one of the λ_α has to be null. The corresponding eigenvector will be, indeed, the stationary distribution, given by $\overrightarrow{u_0} = (P_1^{(eq)}, P_2^{(eq)}, \cdots, P_n^{(eq)})^T$. We can prove this replacing $\overrightarrow{P} = \overrightarrow{u_0}$ in $\partial_t \overrightarrow{P} = \Lambda \overrightarrow{P}$, and considering that $\lambda_0 = 0$. We in fact obtain $\partial_t \overrightarrow{P} = \Lambda \overrightarrow{u_0} = \lambda_0 \overrightarrow{u_0} = 0$, that means it is the equilibrium distribution.
Going back to the non-null eigenvalues, we can add that the smallest of these negative values represents how fast the system converges to the stationary state if we perturb it, and the relative eigenvector is called metastable state and is the state in which the system remains longer.
It is worth noticing that, albeit each eigenvector components decays exponentially with time, the convergence to the stationary distribution can be slower than exponential if a lot of eigenvalues are closer to the metastable one. If the eigenvalues spectrum is closer to an exponential, it can be shown that the practical convergence time is a power law, i.e the convergence mean-time goes to infinity.

Detailed balance The condition of detailed balance is one of the most distinguishing property of a system, and corresponds to the thermodynamic equilibrium. The formal definition regards the microscopic probability flux, asserting the microscopy reversibility: $\lambda_{ij} P_i = \lambda_{ji} P_j$. When this relationship holds, the resolution of the stationary distribution of the master equation is almost trivial for any dimension.
In general this condition can be linked to the Kirchhoff law of fluxes into a network, because the CME can be interpreted as a probability flux on a network generated by the available states and linked by the possible reactions. If we have a circular network of reversible transitions from state A to state B,C and D (so we are working on a system with only 4 possible states) like the graph in fig. 3.2, the condition of detailed balance between fluxes can be rewritten in terms of the reaction propensity alone. Recalling that $P_i = P_j \frac{\lambda_{ji}}{\lambda_{ij}}$, and applying it recursively over the cycle, we can obtain the following condition

$$K_{AB} K_{BC} K_{CD} K_{DA} = K_{DC} K_{CB} K_{BA} K_{AD}. \tag{3.21}$$

In a system that has more than one elementary cycle, the detailed balance condition should be applied to every cycle to be valid for the whole system.

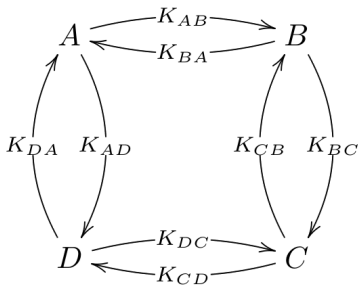


Figure 3.2: Graph with 4 possible states. Figure from [32]

Chapter 4

Ecological theories

In this chapter we are going to describe the main ecological theories proposed to explain the patterns observed in ecological systems, focusing on a simple dynamical model that arises from a Chemical Master Equation approach and that we will apply later to the gut microbiota. At the end of this chapter we will also report some results of a previous application of this dynamical model to ecological systems such as coral reefs.

4.1 Ecological theories purposes and perspectives

The main purpose of modern ecological theories is to describe and explain the within-trophic-level biodiversity [34]. Here, with the term 'biodiversity' we denote both species richness, that is the total number of species in a defined space at a given time, and relative species abundance (RSA), which refers to their commonness or rarity. Instead, with the words 'within-trophic-level' we mean that we are going to study organisms which occupy the same position in a food chain. Thus we will not consider problems such as the trophic organization of communities, or what controls the number of trophic levels, or how biodiversity at one trophic level affects diversity on other trophic levels. The reason for this is that, while not complete, a theory of biodiversity within trophic levels would nevertheless be a major advance because most biodiversity resides within rather than between trophic levels (i.e. there are many more species than trophic levels).
In this perspective, we can define an 'ecological community' as a group of trophically similar species that exist in the same local area and that actually or potentially compete for the same or similar resources, and a 'metacommunity' as the ensemble of all trophically similar individuals and species in a regional collection of 'local communities', in which species may not actually compete because of

separation in space or time.
Modern ecological theories can be distinguished in essentially two main schools of thought: the niche assembly perspective and the dispersal one.
The physicist Heinz Pagels (1982) once observed that there seem to be two kinds of people in the world. There are those who seek deterministic order and meaning in every event, and those who believe events to be influenced, if not dominated, by random chance. This is the controversy between determinism and stochasticity that dominated the twentieth-century physics, one of whose triumphs was exactly to prove that both views of physical nature are simultaneously true and correct, but on very different spatial and temporal scales. The same kind of debate also persists for example in population genetics debates, where the question is whether most change in gene frequencies results from random evolution or from natural selection, and similarly exists in ecology, where there are these two conflicting world views on the nature of ecological communities: the niche and the dispersal perspectives.

Niche Theory The niche assembly perspective holds that communities are groups of interacting species whose presence or absence and even their relative abundance can be deduced from deterministic 'assembly rules' that are based on the ecological niches or functional roles of each species. Here, the concept of 'ecological niche' summarizes the interactions between species and their environment, and is thus defined by two components [47]:

- the requirement for an organism of a given species to live in a given environment (the extent to which a limiting factor, like a resource, a predator or a parasite, influences the birth and death rate of that species);

- the impact of the species on its environment (the extent to which the growth of a population alters the limiting factor, i.e. the availability of a resource or the density of a predator or parasite).

According to this view, species coexist in interactive equilibrium and a stable coexistence among competing species is made possible by niche partitioning. The stability of the community and its resistance to perturbation derive from the adaptive equilibrium of member species, each of which has evolved to be the best competitor in its own ecological niche.
Niche-assembled communities are limited-membership assemblages in which interspecific competition for limited resources and other biotic interactions determine which species are present or absent from the community. We have to under-

line that most proponents of niche assembly come out of a strong neo-Darwinian tradition, which focuses on the lives of interacting individuals and their fitness consequences. The concept of niche follows naturally and logically as the population level summation of the individual adaptations of organisms to their environments.
Niche theory resulted able to predict patterns of species traits and species separation on nutrient gradients similar to those observed in different studies and provided a potential explanation for the high diversity of nature, predicting that habitat heterogeneity can allow a potentially unlimited number of species to coexist if species that are better at dealing with one environmental constraint are necessarily worse at dealing with another [34]. On the other hand, this theory is not able to predict a limit to diversity, and consequently neither to explain species relative abundance.

Dispersal and Neutral Theory The other world view is the dispersal assembly perspective, which asserts that communities are open, nonequilibrium assemblages of species largely thrown together by chance, history, and random dispersal [34]. Species come and go, their presence or absence is dictated by random dispersal and stochastic local extinction.
Actually we will refer to a particular class of dispersal theories, those called 'neutral', in which ecological communities are structured entirely by ecological drift (i.e. demographic stochasticity), random migration, and random speciation. By neutral we mean that the theory treats organisms in a trophically defined community as essentially identical in their per capita probabilities of giving birth, dying, migrating, and speciating (ecological equivalence). We have to underline that neutrality is defined at the individual level, not at the species level, thus this is a very unrestrictive and permissive definition since it does not preclude interesting biology from happening or complex ecological interactions from taking place among individuals. All that is required is that all individuals of every species obey exactly the same rules of ecological engagement. So, for example, if all individuals and species enjoy a frequency-dependent advantage in per capita birth rate when rare, this per capita advantage will be exactly the same for each and every individual of a species of equivalent abundance.
One consequence of a focus on adaptation and niche assembly has been a tendency to accept an equilibrium and a relatively static view of niches and ecological communities. This focus on individual variation in fitness, adaptation and niche, moreover, has led naturally to small-scale, short-term experimental studies of processes of competition, selection and adaptation. Proponents of dispersal assembly criticize this and typically work on much larger spatial and temporal scales, using

biogeographic or paleoecological frames of reference, through an approach less experimental and more analytical of large-scale statistical patterns.
Thus for example, as reported in [34], data from much fossil records revealed that many pre-Holocene, full glacial, and previous interglacial plant communities are very different from modern communities. The evidence from many studies is strong that communities undergo profound compositional changes, sometimes gradual, sometimes episodic, on timescales of centuries to millennia and longer. The fact is that species are transient, even if transit time to extinction are often of the order of millions or ten of millions of year, and furthermore in most of the cases local extinction can not be attributed to competitive exclusion. So, as suggested by Hubbell in his work, we should not concentrate on the indefinite coexistence of specie, but rather on the study of species presence-absence, persistence times, and above all species relative abundance (RSA) in communities, that can be compared with real data.

4.2 Patterns of relative abundance - inductive approaches

Species abundance is of central theoretical and practical importance in conservation biology. In particular, understanding the causes and consequences of rarity is a problem of profound significance because most species are uncommon to rare, and rare species are generally at greater risk to extinction.
Observing the patterns of relative species abundance in different ecological communities (fig.4.1), we can note how all of them have a curiously similar shape, even thought they differ in many ways, including species richness, the degree of dominance of the community by common species, and the number of rare species each community contains. Some are steeper, and some are shallower, but all of the distributions basically exhibit an S-shaped form, bending up at the left end and down at the right end.
Let us now outline the major theoretical and empirical milestones in the study of relative species abundance. First of all we can observe that two major approaches to the study of the distribution of individuals per species have been taken: inductive and deductive. In the early years, when the study of relative species abundance was in its infancy, the inductive approach dominated. Observed distributions of the numbers of individuals per species in collections were fit to statistical distributions with little or no attempt to give a theoretical explanation or to define the sampling universes from which the collections were made.

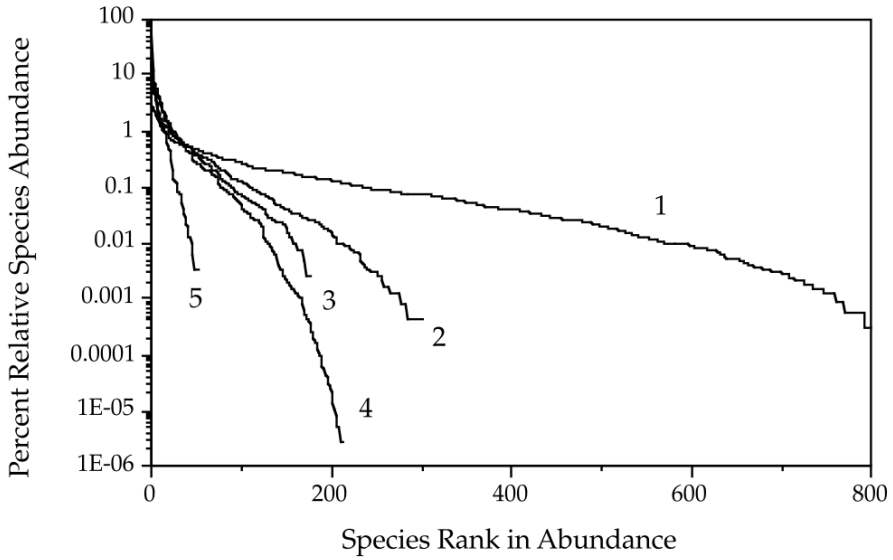


Figure 4.1: Patterns of relative species abundance in a diverse array of ecological communities from [34]. Species in each community are ranked in percentage relative abundance from commonest (left) to rarest (right). The percentage relative abundance is log transformed on the y-axis. (1) Tropical wet forest in Amazonia. (2) Tropical dry deciduous forest in Costa Rica. (3) Marine planktonic copepod community from the North Pacific gyre. (4) Terrestrial breeding birds of Britian. (5) Tropical bat community from Panama.

Fisher's logseries distribution A milestone of the inductive approach was the work proposed by Fisher, Cobert and Williams in 1943 [30].
Corbet and Williams, studied abundance data, respectively about butterflies in Malaya and moths collected over a four-year period at the Rothamsted Experimental Station in England, and plotted the number of species into abundance classes, i.e. species represented by a single individual, by two individuals, and so on. In these plots, the authors both noticed that the series was a relatively smooth hyperbolic progression, with many rare and few common species.
When then Fisher analyzed their data, he assumed that relative abundances of species in nature would be well described by a gamma function and that the number of individuals collected of a given species would be Poisson distributed because most species were rare and represented by only a few individuals in the samples of Corbet and Williams. The resulting compound distribution was negative binomial. However, there was a problem because the zero abundance class (species too rare to be sampled) was obviously not observable, so Fisher truncated the negative binomial to eliminate the zero class. Then, having no way of estimat-

ing how many species were not sampled, Fisher assumed the number of species in the community was effectively infinite. Fisher obtained a one-parameter distribution that he dubbed the logarithmic series, derived from the negative binomial as a limiting case (shape parameter set to zero).
According to the logseries, as it is now generally called, the number of species in a collection having n individuals will be given by

$$\alpha x^n/n, \tag{4.1}$$

where x is a positive constant $0 < x < 1$ and α is a measure of diversity, which in the expectation is equal to the number of singleton species divided by x. Thus, the number of species with $1, 2, 3, 4, \ldots n$ individuals will be given by $\alpha x, \alpha x^2/2, \alpha x^3/3, \alpha x^4/4, \ldots, \alpha x^n/n$ for $0 < x < 1$.
Adding all terms, the total number of species, S, is expected to be $\alpha[-ln(1-x)]$, and the total number of individuals in the collection, N , is $\alpha x/(1-x)$. The parameter α, known as Fisher's α, is a widely used measure of species diversity because it is theoretically independent of sample size [30], even if in other studies [34] is showed that empirically α is only approximately constant, changing slowly over large ranges in sample size.
Fitting the logseries always results in the singleton category having the most species, as shown in fig.4.2.

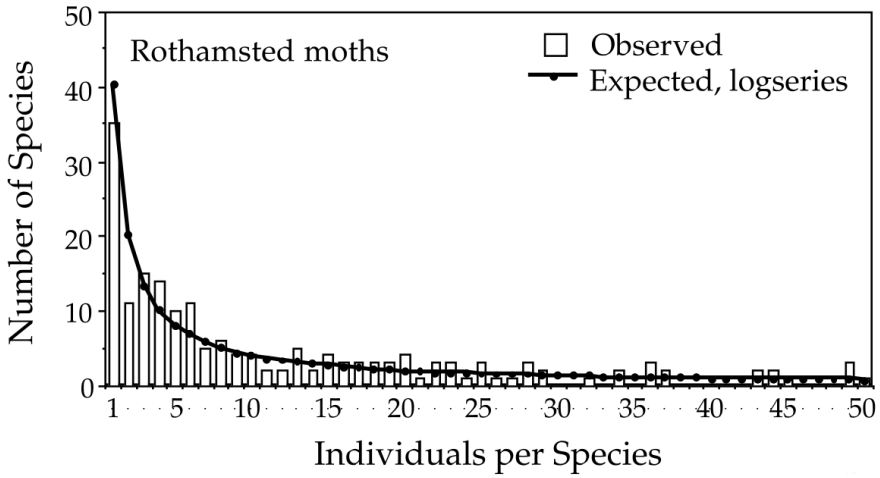


Figure 4.2: An example from [34] of the use of the logseries distribution to fit data on species abundance in collections of months.

Preston's lognormal distribution A few years later, Preston (1948) criticized the logseries on the grounds that it was not a good fit to data that he had assembled, primarily on bird species abundances. Preston argued that relative abundance distributions were often bell-shaped curves, such that species having intermediate abundances were more frequent than very rare species. Preston actually noted that the distributions were lognormal and introduced a simple way to display this lognormal distribution of relative species abundance. He built doubling categories of abundance (1, 2, 4, 8, etc.), and counted the species having abundances falling in each category. Species having exactly 1 2 4 8 individuals were divided equally between adjacent abundance categories. He called these doubling classes 'octaves' in analogy to octaves of a musical scale, which represent a doubling of the frequency of musical pitch. This classification of species into doubling abundance classes effectively log transforms the relative abundance data to the log base 2. He chose log base 2 for the simple practical expedient of spreading the distribution of species abundances over more categories to make its shape more apparent. Using any larger number for the base of log transforming the distribution would only reduce the number of categories displayed, depending on the range in relative species abundances.

The lognormal distribution is continuous, not discrete as in the case of the logseries. However, Preston's method of categorizing abundances provides a simple way to approximate the distribution by a discrete-valued function, as follows. Let S_0 be the number of species in the modal octave of abundance. Let S_R be the number of species in the $R-$th octave (or doubling abundance class) to the left or right of the modal octave. Then the so called Species Curve can be written as

$$S_R = S_0 e^{-a^2 R^2}, \tag{4.2}$$

with $R = 0, 1, 2, \ldots$ and where a is a constant that depends on the variance of the lognormal, $a = 1/\sqrt{2\sigma}$. Note that the distribution is symmetrical about the mode, located at $R = 0$. Fitting the Species Curve can be done approximately by taking natural logs, and regressing $ln(S_R)$ on R^2, a regression having slope $-a^2$ and intercept $ln(S_0)$. More accurate fitting of the continuous lognormal distribution to the data on individual species abundances requires using a maximum likelihood technique for a truncated lognormal.

Over the past half century, the lognormal distribution has been fit successfully to a far larger number of relative species abundance distributions than has the logseries distribution, particularly as larger sample sizes have become available [34].

To explain his lognormal distribution, Preston argued that the shape of the relative species abundance distribution observed by Fisher and his colleagues was an artifact of small sample size. In the logseries, the expected number of species is always largest in the rarest abundance category, consisting of singleton species. However, in a small sample, one should observe only a truncated distribution of

relative abundances, comprising only the most common species. This is because common species are generally collected sooner than rare species.

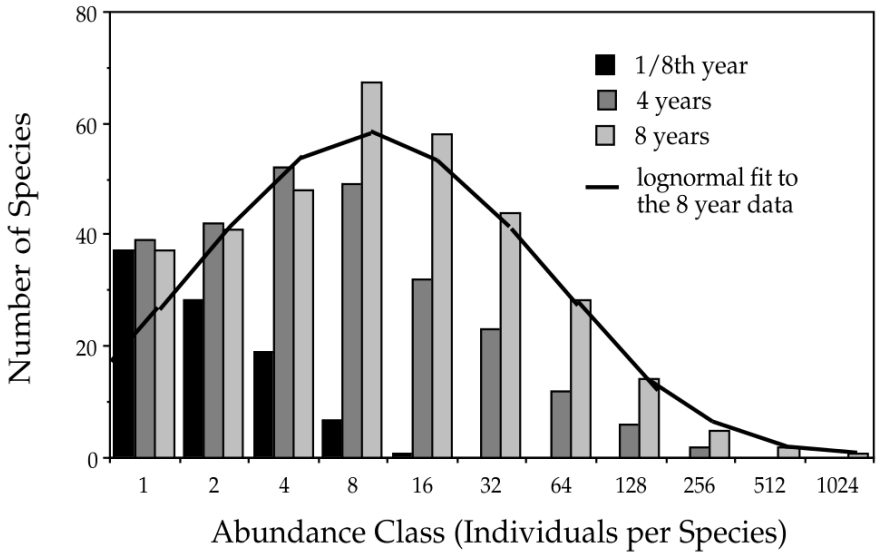


Figure 4.3: As the survey of months at light traps at Rothamsted Field Station was extended over more years, the distribution of individuals per species became lognormal, as Preston predicted. Figure from [34].

As sample size increases, Preston predicted that more and more of the lognormal distribution would be revealed, as shown in fig.4.3, and the reason for which Fisher had not noted it was because they did not consider the importance of sample size, because of the theoretically expected constancy of Fishers α in collections of different sizes.
However, the proposal for a lognormal distribution let the apparent invariance of Fisher's α unexplained and, furthermore, in recent years, as larger sample sizes of relative species abundance have become available and the abundances of very rare species have become better known, it has become increasingly apparent that observed distributions of relative species abundance are actually seldom lognormally distributed. Observed distributions appear to be lognormal to the right of the mode in the right-hand tail representing common species. But they almost always show a strong negative skewness, as we can see in fig.4.4, that can not be explain neither with Fisher or with Preston's distribution.

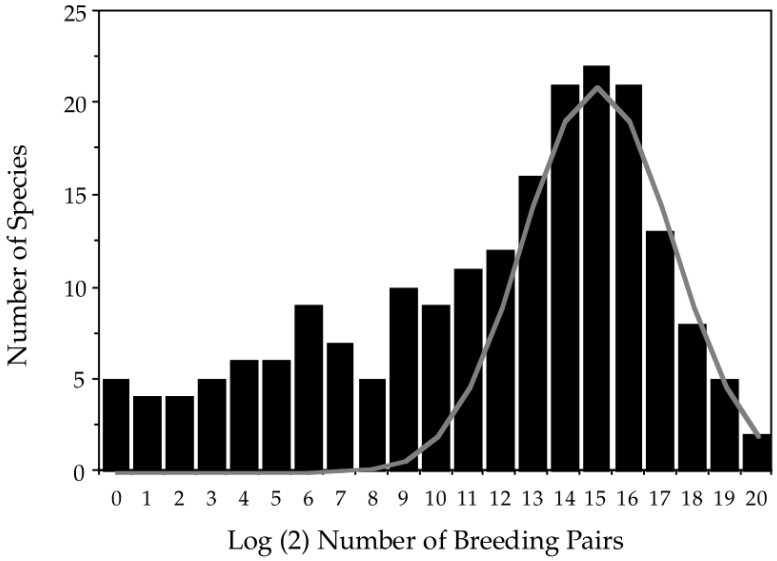


Figure 4.4: Negatively skewed distribution of relative species abundance for all British breeding birds. Note the poor fit of the lognormal to the left-hand tail of rare and extremely rare species. Figure from [34].

4.3 Patterns of relative abundance - deductive approaches

Early deductive approaches As we said before, the logseries and lognormal were the principal inductive approaches to the study of relative species abundance. These were followed by different attempt to derive a theory of relative species abundance from first principles, that is based on hypotheses about how ecological communities were organized. You can find a review of these early approaches in [34]. Here we just note that these early theories present many shortcomings, among which a not entirely exhaustive biological interpretation, the presence of many parameters, and, besides that, the inability to explain real data.
Furthermore, all these early approaches belong to the niche perspective and consequently they tend to accept the idea of coexistence as a static equilibrium and to describe the ecological communities through a static point of view. None of them in fact tries to undertake a dynamical approach and to make the theory emerge from natural processes of birth, death and dispersion, and this is of course due to their focusing on coexistence rather than on relative abundance distribution, following Lotka-Volterra's tradition.

MacArthur and Wilson theory of island biogeography The first deductive theory for ecological system, based on the idea of a dynamical equilibrium, and that introduced the concept of neutrality in ecology, is that built by MacArthur and Wilson in their monography of 1967 [37].
MacArthur and Wilson erected their theory in part to explain the puzzling observation that islands nearly always have fewer species than areas on continents of the same size, a fact that could not been explained with the current static idea of island communities, that considered them fixed over ecological time-scale ($\sim 10^3$ years). The authors proposed a new theory, in which the number of species present on the island changes as a result of two opposing forces: immigration from the continents of species not already present on the island, and extinction of species present on the island.
MacArthur and Wilson reasoned that this possibility of extinction comes from the fact that island average population sizes are smaller than those on the continents, and small populations are subject to some complications:

- the Allee effect, that is at small densities the death rate could be higher than the birth one because of difficulties in finding a partner for reproduction, higher exposition to predators, disruption of the social structure of a population;
- the inbreeding, that is the coupling between related individuals, that brings to the genetic deterioration of the population, that is to a reduced survival and fertility;
- the strong influence of adverse casual events: the probability that few individuals die at the same time, is higher that the probability that many individuals die all together (demographic stochasticity).

Furthermore, once island populations went extinct, it would take the same species longer to recolonize the island than it would take them to disperse among adjacent areas on the mainland. Thus, other things being equal, species would spend a smaller fraction of total time resident on a given island than in the same-sized area of the mainland. Given these assumptions, i.e., a higher island extinction rate and a lower reimmigration rate, one then predicts a lower steady-state number of species on islands than in same-sized areas on the mainland. MacArthur and Wilson captured this simple equilibrium idea in a now famous graph (fig.4.5).
The equation that controls the species dynamics is

$$\frac{dS}{dt} = I(S) - E(S), \tag{4.3}$$

where $I(S)$ is the immigration rate of new species, while $E(S)$ is the extinction rate. $I(S)$ will be a descending function, as the probability of immigration of a

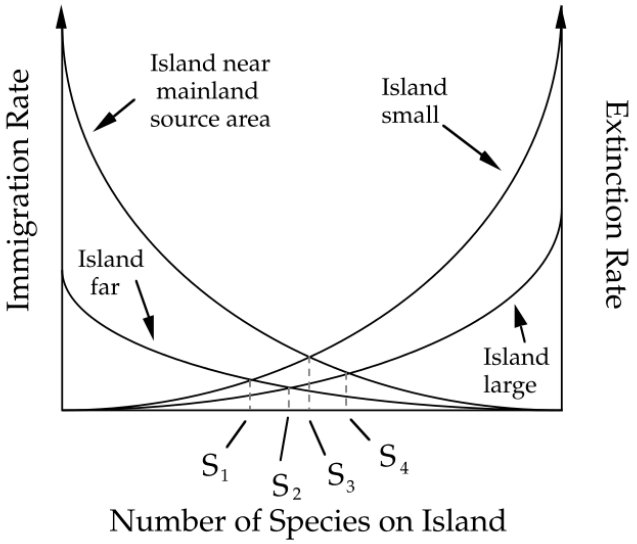


Figure 4.5: MacArthur and Wilson's equilibrium hypothesis for explaining the number of species on islands (S_i) as a dynamic equilibrium between the rate of immigration of new species onto the island and the rate of extinction of species already resident on the island. Figure from [34].

new species will be higher if the species already present on the island are few. For what concern $E(S)$, instead, this will be an ascending function, that is null for $S = 0$. In fact, we can write $E(S) = \mu(S)S$, where μ is the probability that a single species get extinct in unit of time. $\mu(S)$ will be an ascending function of the number of species on the island, in fact if there are many species, there will be a higher interspecific competition, that is a lower number of free niches on the island, and this will increase the probability of a species to get extinct.
The stationary state will be that in which the immigration and the extinction rate become equal, that is the point S_1, S_2, S_3 or S_4 in fig. 4.5, that refer to different conditions in which we can find the island.
In its fundamental assumptions MacArthur and Wilson theory is a dispersal theory that asserts that island communities are dispersal assembled, not niche assembled. It predicts a steady-state number of species on islands under a persistent rain of immigrant species from mainland source areas. However, in contrast to niche-assembly community theory, it predicts only a diversity equilibrium, not a taxonomic equilibrium: what remains steady here is the number of species on the island, and not their identity, in fact in the steady point the immigration and the extinction rate are equal but not null. Consequently, there is a turnover of the present species, that is exactly $I(S_i) = E(S_i)$ for unit of time. Thus, we can state

that we are dealing with a theory that can be considered neutral, since all species are considered equal in their probabilities of immigration onto the island, or of going extinct once there.
Breaking away from the conventional neo-Darwinian view, ecological communities are now described as in perpetual taxonomic nonequilibrium, undergoing continual endogenous change and species turnover through repeated immigrations and local extinctions (let us note that anyway it does not include speciation). These turnovers need not be especially rapid, however, and species can coexist for long periods in slowly drifting mixtures and in shifting relative abundances. Thus, MacArthur and Wilson theory highlights the possibility that environmental and demographic stochasticity can play a role equal or also more important than niche assembly rule in structuring ecological communities.

Caswell's abundance random walk In the mid-1970s, when most eyes were still focused on the classical, niche-based theory of community ecology, Caswell made a bold attempt to create a neutral theory of community organization. Borrowing mathematical machinery from the theory of neutral evolution in population genetics, Caswell erected his model, in which communities are essentially collections of completely noninteracting species in which each species undergoes an independent random walk in abundance. Therefore, the total size of the community fluctuates.
New species enter the community as a Poisson process (i.e., a rare event) with probability ν per unit time. This immigration probability, as in the theory of island biogeography, is independent of the identity of the species and of the number and identities of the species already present, except that only species not currently present are allowed to immigrate. This is equivalent to assuming that immigration makes a negligible contribution to the population dynamics of a species already present. Each new immigrant species becomes the founder of a line of descendants.
Caswell assumed a linear birth-death process in which the stochastic per capita birth and death rates, λ and μ, are assumed to be equal, corresponding to the deterministic case of a zero intrinsic rate of increase, r. In other words, each species population is as likely to increase as it is to decrease per unit time. This is a pure drift process or random walk. The transition probabilities from a population of size N_i to size N_{i-1}, N_i , or N_{i+1} at time $t + dt$ are linear functions N_i of at time

t, as follows:

$$\begin{aligned} Pr(N_{i-1}|N_i) &= \mu N_i \\ Pr(N_i|N_i) &= 1-(\lambda+\mu)N_i \\ Pr(N_{i+1}|N_i) &= \lambda N_i \end{aligned} \tag{4.4}$$

Note in this model that λ and μ must be chosen to be sufficiently small to satisfy $(\lambda+\mu)N_t < 1$.
Actually Caswells model has different problems. First of all its results differ substantially from observed community relative abundance patterns and look decidedly not lognormal on a Preston plot of octaves of abundance.
Furthermore, there are also more serious problems with Caswell's model. One is that the size of the community grows without bound over time. Community size, J , where J is the total number of individuals in the community, is a negative binomial random variable with mean $E[J] = t \to \infty$ (elapsed time), and variance $Var[J] = t(t+1) \to \infty$ as $t \to \infty$. A second major problem is that the expected number of species in the community, $E[S]$, is linearly proportional to the colonization rate of new species per unit time, ν, and the log of elapsed time:

$$E[S] = Var[S] = \nu \cdot ln(t+1). \tag{4.5}$$

Despite all these defects, this model is very important since it was the first model of relative species abundance explicitly based on birth, death, and dispersal processes. Moreover, with the addition of the assumption of a finite community size (due to limited resource availability) and minor changes in the birth, death, and dispersal processes, a much better model can be obtained.

4.4 Dynamical models of RSA

As we have seen in previous sections, many niche and neutral models have been proposed to explain the RSA trend observed in so many experimental data.
We now present a simple unified theory for understanding these RSA patterns, developed by Azaele et al. [21] in the continuous form and by Volkov et al. [49] in the discrete one.
This theory offers an explanation for diversity, species composition, relative species abundance patterns, and invasion dynamics in ecological communities, resolving many of the shortcomings of both classical niche theory and neutral theory.
The basis of this theory is that niche partitioning and demographic stochasticity are both involved in structuring communities. Such combined approaches might offer an explanation for the diversity, composition and relative abundance patterns

of species observed in ecological communities.

Dynamical Model - continuous form We treat the population of a species at time t as a continuous variable, $x(t)$, an assumption which is valid when the population varies smoothly with time and is not too small.
We assume that the species population is subject to two distinct dynamical processes. The first of these is deterministic whereas the second one is stochastic. The deterministic process has two contributions: 1) an immigration rate b, which, for simplicity, is assumed to be equal for all species and independent of time and 2) an effective competition term proportional to the population of the species which serves to fix the average population.
The stochastic process controls the demographic fluctuations, not considered by the deterministic part, and is proportional to $\sqrt{x}$ as described in appendix A.
We know that a system ruled by a deterministic component, described by a vectorial field $a(x,t)$, and by some white noise $b(x,t)\xi(t)$, that reflects a stochastic component, can be modeled by the Langevin equation

$$\frac{dx}{dt} = a(x) + b(x)\xi(t), \tag{4.6}$$

and that the equation for the probability density function corresponding to this process is the Fokker Planck equation

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}(\rho a(x)) + \frac{b(x)^2}{2}\frac{\partial^2 \rho}{\partial x^2} \tag{4.7}$$

as demonstrated in [43].
Thus, the Langevin equation corresponding to our system is

$$\dot{x}(t) = b - \frac{x(t)}{\tau} + \sqrt{Dx(t)}\xi(t), \tag{4.8}$$

where $x > 0$ for any $t > 0$; b, τ and D are positive real constants, $\xi(t)$ is a Gaussian white noise, that means it has zero mean value and time correlation $\langle \xi(t)\xi(t')\rangle = 2\delta(t-t')$.
The corrisponding Fokker Planck equation for this process is

$$\dot{p} = \partial_x[(x/\tau - b)p] + D\partial_x^2(xp), \tag{4.9}$$

where $p = p(x,t)$ is the probability density function (pdf) of finding x individuals at time t in the community, i.e. $\int_n^{n+\Delta n} p(x,t)dx$ is the fraction of species with

a population between n and $n + \Delta n$. Setting $\dot{p} = 0$ one obtains the stationary solution of eq.4.9

$$p_0(x) = (D\tau)^{-b/D}\Gamma(b/D)^{-1}x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}}, \tag{4.10}$$

where $\Gamma(x)$ is the gamma function.
In fact $\dot{p} = 0 \Rightarrow \partial_x[(x/\tau - b)p] + D\partial_x^2(xp) = 0$ that means

$$\partial_x\left([(x/\tau - b)p] + D\partial_x(xp)\right) = 0. \tag{4.11}$$

The obvious solution is $[(x/\tau - b)p] + D\partial_x(xp) = constant$ and we can set $cost = 0$ obtaining

$$D\partial_x(xp) = (b - x/\tau)p. \tag{4.12}$$

We can rewrite the equation in the form $\partial_x g = A(x)g(x)$ multipling and dividing the right hand side by Dx

$$\partial_x(Dxp) = \frac{(b - x/\tau)}{Dx}Dxp \tag{4.13}$$

where $g(x) = Dxp$ and $A(x) = \frac{(b-x/\tau)}{Dx}$. Thus the solution will be $g(x) = e^{\int A(x)dx}$, i.e.

$$\begin{aligned} p_0(x) &= \frac{1}{Dx}e^{\int \frac{b-x/\tau}{Dx}dx} = \frac{1}{Dx}e^{\frac{b}{D}logx - \frac{1}{D\tau}x} \\ &= \frac{1}{D}x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}} \end{aligned} \tag{4.14}$$

Then we have to normalize this function to finally find the stationary solution. Thus we calculate

$$\frac{1}{D}\int_0^\infty x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}} = 1, \tag{4.15}$$

that we can rewrite, multiplying and dividing for $(D\tau)^{\frac{b}{D}-1}$, as

$$\frac{1}{D}(D\tau)^{\frac{b}{D}-1}\int_0^\infty \frac{x^{\frac{b}{D}-1}}{(D\tau)^{\frac{b}{D}-1}}e^{-\frac{x}{D\tau}} = 1. \tag{4.16}$$

Now we can make the change of variables $\frac{x}{D\tau} = t$ and $\frac{dx}{D\tau} = dt$, and the equation becomes

$$\frac{1}{D}(D\tau)^{\frac{b}{D}-1}\int_0^\infty t^{\frac{b}{D}-1}e^{-t}D\tau dt = 1. \tag{4.17}$$

Now if we introduce the definition of the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, our equation can be written as

$$\frac{1}{D}(D\tau)^{b/D}\Gamma(b/D) = 1 \Rightarrow \frac{1}{D} = \frac{(D\tau)^{-b/D}}{\Gamma(b/D)}, \tag{4.18}$$

and the stationary solution becomes

$$p_0(x) = P_{RSA} = \frac{(D\tau)^{-b/D}}{\Gamma(b/D)} x^{\frac{b}{D}-1} e^{-\frac{x}{D\tau}} \tag{4.19}$$

This solution obeys reflecting boundary conditions at $x = 0$ which, in a stationary regime, fix the number of species on average. Thus the steady-state solution $p_0(x) = P_{RSA}(x)$, which is independent of initial conditions, provides an exact expression for the relative species abundance (RSA). Furthermore, for $b/D \ll 1$, one obtains the RSA distribution of metacommunity proposed by Fisher.
In order to understand the meaning of the parameters involved in the equation 4.10, we can start from a discrete master equation. The time evolution of a population x of a given species is governed by the following birth-death master equation ($x = 1, 2, \cdots$)

$$\frac{\partial p_x(t)}{\partial t} = b(x-1)p_{x-1}(t) + d(x+1)p_{x+1}(t) - [b(x) + d(x)]p_x(t) \tag{4.20}$$

where the birth and death rates are given by

$$\begin{aligned} b(x) &= x(b_1 + b_0/x) \\ d(x) &= x(d_1 + d_0/x) \end{aligned} \tag{4.21}$$

Here, b_1 and d_1 are the per-capita rates and the presence of the constants b_0 and d_0 produce a density dependence effect, which causes a rare species advantage (disadvantage) when $b_0 > d_0$ ($b_0 < d_0$). Such density dependence can arise due to effective rates of immigration/emmigration/speciation/extinction in a local community. The skewness of the RSA indicates a rare species advantage and thus $b_0 > d_0$ and for simplicity and parsimony we will choose $b_0 = -d_0$ in the following. Treating x as a continuous variable we use the Taylor expansion

$$d(x+1)p_{x+1}(t) - d(x)p_x(t) = \frac{\partial}{\partial x}(d(x)p_x(t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(d(x)p_x(t)) + ... \tag{4.22}$$

and similarly for $b(x-1)p_{x-1}(t) - b(x)p_x(t)$. Thus the previous master equation becomes a Fokker-Planck (FP) equation, whose explicit form is

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[d(x) - b(x)]p(x,t) + \frac{1}{2}\frac{\partial^2}{\partial x^2}[d(x) + b(x)]p(x,t). \tag{4.23}$$

By inserting the rates in 4.21 into this FP equation and setting $b_0 = -d_0 > 0$, we obtained

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[(d_1 - b_1)x - b]p(x,t) + \frac{[d_1 + b_1]}{2}\frac{\partial^2}{\partial x^2}xp(x,t). \tag{4.24}$$

This equation matches with 4.9 on setting

$$\begin{aligned} D &= \frac{d_1 + b_1}{2} \\ \tau &= \frac{1}{d_1 - b_1} > 0 \\ b &= 2b_0 > 0 \end{aligned} \tag{4.25}$$

that are three parameters with ecological significance: D describes the effects of demographic stochasticity (fluctuations) and is given by $(b_1+d_1)/2$; $\frac{1}{\tau} = d_1 - b_1$ is the imbalance between the death and birth rates that inexorably drives the ecosystem to extinction in the absence of immigration or speciation, thus τ is the characteristic time associated to species turnover in neutral evolution (an ecosystem close to the stationary state is able to recover from a perturbation on a timescale of order τ); $b = 2b_0$ describes the density dependence effects arising from immigration and/or speciation.
Let us observe that in the more general case, with $b_0 \neq -d_0$, one obtains an extra term in the FP equation, i.e. Dx is substituted by $Dx + (b_0 + d_0)/2$. However, we can neglect the new term if $Dx \gg (b_0 + d_0)/2$, in fact when $|b_0| \sim |d_0|$ is of the order of $b/2$, as found in different real data, $(b_0 + d_0)/2$ is at most $D/4$. Anyhow, the same reasoning for the term $x/\tau - b$ does not allow one to neglect the b term because τ is large.
Thus the FP equation with the three parameters we have used is consistent even when $b_0 \neq -d_0$.

Dynamical Model - discrete form Following the work of Volkov et al. [49], we can also find the discrete form of the RSA distribution, considering a birth-death process for a discrete community.
In this case we will have a discrete variable n that describes the number of individuals of a k-th species, so that $P_{n,k}(t)$ will denote the probability that the k-th species contains n individuals at time t. Thus, now we have that the time evolution of $P_{n,k}(t)$ is regulated by the master equation

$$\frac{\partial P_{n,k}(t)}{\partial t} = P_{n-1,k}(t)b_{n-1,k} + P_{n+1,k}(t)d_{n+1,k} - P_{n,k}(b_{n,k} + d_{n,k}), \tag{4.26}$$

that is equivalent to equation 4.20 with birth rate $b_{n,k}$ and death rate $d_{n,k}$ related to the k-th species with n individuals.
We can solve this equation to find the stationary solution using the linear expansion method. This method exploits the fact that in a birth-death process we can

state that in stationary conditions, the flux should be null at each step, thus

$$P_{n,k}b_{n,k} = P_{n+1,k}d_{n+1,k}, \tag{4.27}$$

that is the detailed balance condition (see paragraph 3.3).
This allows us to write a simple recursive solution for this system

$$P_{n+1,k} = P_{n,k}\frac{b_{n,k}}{d_{n+1,k}}. \tag{4.28}$$

and so, given a certain value of $P_{0,k}$ to maintain the final normalization of probability, we can write the stationary solution that is reached in the infinite time limit, as

$$P_{n,k} = P_{0,k}\prod_{i=0}^{n-1}\frac{b_{i,k}}{d_{i+1,k}}. \tag{4.29}$$

Let us consider a simple, ecologically meaningful form for the effective birth and death rates of the k-th species

$$\begin{aligned} b_{n,k} &= b_k(n + \Upsilon_k) \\ d_{n,k} &= d_k n \end{aligned} \tag{4.30}$$

where b_k and d_k denote the per-capita density-independent birth and death rates and a non-zero Υ_k could arise from either immigration or owing to intraspecific interactions such as those giving rise to density dependence. We do not incorporate speciation explicitly into the model because it does not affect the functional form of the results (it can be incorporated into the immigration term at $n = 0$ by adding a constant); we do not even explicitly include emigration since it depends by n and can thus be considered already expressed in the term $d_k n$.
The steady-state solution of the master equation for $P_k(n)$, the probability that the k-th species has n individuals, yields a negative binomial distribution

$$P_{n,k} = P_{0,k}\frac{b_k^n}{d_k^n n!}\Gamma(n + \Upsilon_k) = \frac{(1 - x_k)^{\Upsilon_k}}{\Gamma(\Upsilon_k)}\frac{x_k^n}{n!}\Gamma(n + \Upsilon_k), \tag{4.31}$$

where $x_k = b_k/d_k$, the ratio of the per-capita birth rate to the per-capita death rate, controls the mean species abundance given by $x_k\Upsilon_k/(1 - x_k)$, and where $P_{0,k}$ was deduced by the normalization condition $\sum_n P_{k,n} = 1$. Furthermore, as we showed in 3.3, the system always reaches the stationary condition for $t \to \infty$.
The number of species containing n individual is given by

$$\varphi_n = \sum_{k=1}^{S} I_{n,k}, \tag{4.32}$$

where S is the total number of species that may potentially be present in the community and $I_{n,k}$ is a random variable that takes the value 1 with probability $P_{n,k}$ and 0 with probability $(1 - P_{n,k})$. Thus the average number of species containing n individuals is

$$< \varphi_n >= \sum_{k=1}^{S} I_{n,k} P_{n,k} = \sum_{k=1}^{S} P_{n,k}, \tag{4.33}$$

from which follows the condition

$$\sum_{n} < \varphi_n >= S. \tag{4.34}$$

Let us now suppose that the probabilities of birth and death for each individual do not depend on the species to which it belongs. Thus, let us impose that the species in the metacommunity are demographically equal, that is we are introducing the hypothesis of neutrality in our system

$$\begin{aligned} b_k &= b \\ d_k &= d \\ x_k &= x. \end{aligned} \tag{4.35}$$

From equation 4.33 and 4.31, we finally have that

$$< \varphi_n >= \frac{S}{[(1-x)^{-\Upsilon} - 1]\Gamma(\Upsilon)} \frac{x^n}{n!} \Gamma(n + \Upsilon), \tag{4.36}$$

and placing $\theta = S/[(1-x)^{-\Upsilon} - 1]\Gamma(\Upsilon)$, that is Fisher biodiversity number, an estimate of species variety inside our ecosystem, we have

$$< \varphi_n >= \theta \frac{x^n}{n!} \Gamma(n + \Upsilon). \tag{4.37}$$

We can observe that the biodiversity number θ does not depend from the number of individuals in the community, but, as required by Fisher, only from the rate $x = b/d$. Let us remark that in these dynamical models, we used two key assumption: we considered non-interacting species and individuals neutrality. With these two assumptions we were able to obtain a deductive model in which Fisher distribution and its biodiversity parameter emerge as a consequence of the birth-death processes of the individuals in the species.

4.5 Application: coral reefs

To show an application of this dynamical model, we refer to the work of Volkov et al. [49], in which the authors described the RSA patterns of coral reefs through

the model reported in section 4.4.
The starting point for this work was that previous studies on coral reefs patterns reported log-series-like RSA distributions in local communities, and log-normal-like RSA distributions when a geographically widespread set of coral-reef communities was pooled to estimate the RSA distribution for the metacommunity. In their study, Volkov et al. considered a relatively small semi-isolated local community surrounded by a very large metacommunity acting as a source of immigrants. In coral reefs, in fact, each local community receives immigrants from all the surrounding semi-isolated local communities, within each of which the species abundances are not frozen in time (as it is instead for rainforests). Let us observe that we can suppose we have the same conditions in the gut microbiota communities that we will analyze.
In their work, Volkov et al. fitted the coral-reef RSA distribution using equation 4.37, as shown in figure 4.6.

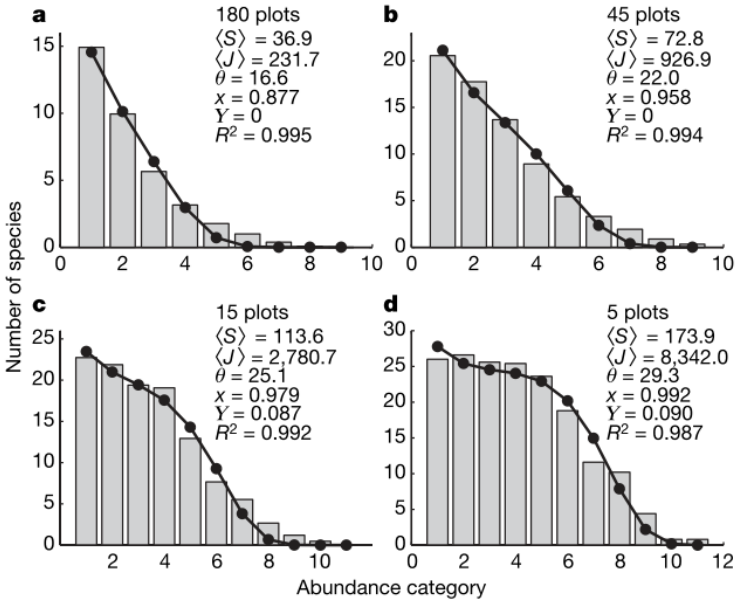


Figure 4.6: Preston plot and fits of equation 4.37 to the coral-reef species abundance data for: 180 coral-reef local communities (a), 45 reef communities, each of which consisting of 4 local communities (b), 15 metacommunities, each of which consistng of 5 reef communities (c), and metacommunity (habitats pooled), each of which consisting of 5 metacommunities (d). Figure from [49].

To explain these RSA distributions, Volkov et al. applied the model reported in section 4.4. First of all, they considered that, given the isolation of individual

coral reefs under the island metacommunity model, the value of the immigration parameter Υ is very small since local communities are separated from each other by large distances. In such a situation, the RSA for the local communities resembles the Fisher log-series, and does not have an interior mode (at abundance $n > 1$). Then they tried to gradually assemble the metacommunity RSA distribution by considering the joint RSA distributions of multiple local communities in the following way. First consider the joint RSA of two local communities A and B making up the metacommunity. Consider a species that has n_A individuals in community A with probability $P(n_A)$ and n_B individuals in community B with probability $P(n_A)$. The probability that the species has n individuals in A and B is demonstrated to be

$$P(n_A + n_B = n) = \sum_{n_A+n_b=n} P(n_A)P(n_B) \propto \frac{x^n}{n!}\Gamma(n + 2\Upsilon). \tag{4.38}$$

The corresponding RSA has the same form as for the single community but with the effective immigration parameter 2Υ. Extending the calculation of the joint RSA distribution to more and more local communities, one arrives at the RSA of the metacommunity characterized by an effective immigration parameter $L\Upsilon$, where L is the total number of local communities making up the metacommunity. When L is large, the RSA distribution exhibits a clear, interior mode at abundance $n > 1$, and the rare species constitute a smaller fraction of all the species than in the local community. We can in fact see from the fits that, on local scales (local communities and reefs), immigration is almost absent so that the local RSA resembles the Fisher log-series distribution. Their theory thus explained how the RSA becomes log-normal-like on aggregating the local communities into one metacommunity.

Let us finally observe that this mean field analysis does not take into account the actual spatial locations of the local reef communities. This fact is important since we are going to apply the same model to a gut microbiota community, sampled by sequencing, in which of course we do not have spatial informations.

Chapter 5

Results

In this chapter we report our results on gut microbiota next-generation sequencing data analysis. First of all we will describe our dataset, then we will explain how, through a clustering procedure, we were able to deduce the bacteria species distribution and to fit it with the gamma distribution theoretically obtained in chapter 4. Finally we will compare our results, showed in the form of Preston plots, with the coral reefs study reported previously in section 4.5.

5.1 Data

In order to analyze the Relative Species Abundance (RSA) distributions, as a measurement of the gut microbiota biodiversity, we downloaded from the SRA (Sequence Read Archive) database of NCBI [7], data from the experiment of Jeraldo et al. [35].
Jeraldo's data are composed of five samples, as described in table 5.1.

Sample	Description	Treatment	Spots	Bases	Size
SRR491179	chicken cecum	inoculated with C. jejuni 1 wk before cecal sampling	18324	3.8M	8.9Mb
SRR491180	cattle rumen 1	sampled at 0 h after feeding (gut sample)	47489	7.7M	4Mb
SRR491181	cattle rumen 2	sampled at 8 h after feeding (gut sample)	31074	5.2M	2.5Mb
SRR491182	swine clone 1	fed with diet 1 before fecal sampling	42443	9M	19.5Mb
SRR491183	swine clone 2	fed with diet 2 before fecal sampling	44927	9.5M	20.3Mb

Table 5.1: Description of Jeraldo's data.

16S rRNA data of these samples were obtained using 454 Life Sciences pyrose-

quencing.

5.2 Clustering

As explained in section 1.3.2, our aim was to cluster 16s rRNA sequences into OTUs (Operational Taxonomic Units, correspond to bacteria taxa at a particular taxonomic level), in order to quantify their respective abundance and to compute the relative abundance distribution of bacteria species in one sample's gut microbiota.
For this purpose, after converting the .sra files to .fastq and then to .fasta, we computed a *de novo* clustering (that is without a reference database) with UCLUST (see section 1.3.2), using different similarity thresholds.
The similarity thresholds used for UCLUST clustering are: 97%, 90%, 85%, 80%, 75%, 70%, and 65%. Let us underline that, with the decreasing of the similarity threshold, we will obtain more abundant OTUs and, at the same time, the total number of OTUs generated will decrease.
As we underlined in section 1.3.2, before clustering with UCLUST it is recommended to sort the input sequences so that the seeds would be chosen among the best. Since our data are obtained with 454 Life Sciences pyrosequencing, a preferable way to sort them is by quality score.
The quality score is a code the is provided in the .fastq file, and is assigned to each nucleotide base call in automated sequencer traces. The quality score has become, by this time, widely accepted to characterize the quality of DNA sequences.
454 quality scores are expressed as Phred scores Q [11], which are defined as a property which is logarithmically related to the base-calling error probabilities P:

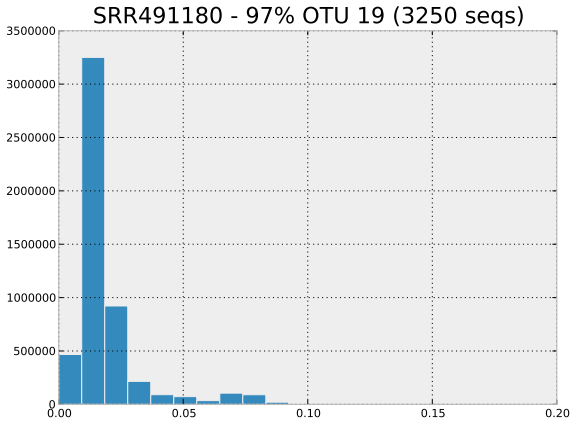
$$Q = -10 \cdot log_{10} P. \tag{5.1}$$

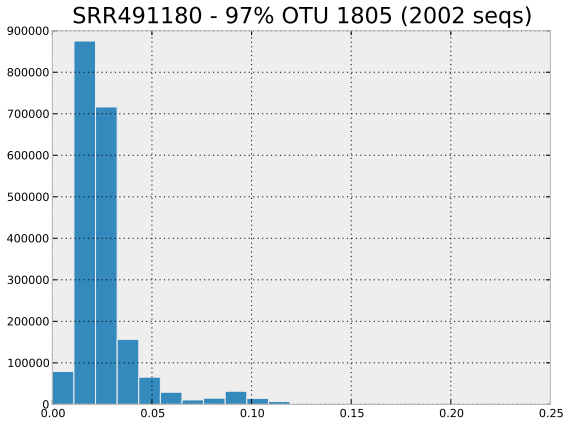
For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000, while the base call accuracy is 99.9%. If instead the quality score is 10, the probabilities that the base is incorrectly called are 1 in 10 and the base call accuracy is 90%. Thus, we ordered our input sequences by quality score, putting first those with higher quality score.

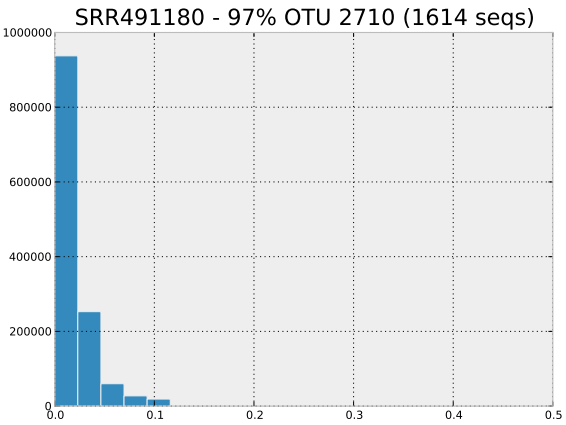
5.3 UCLUST tests

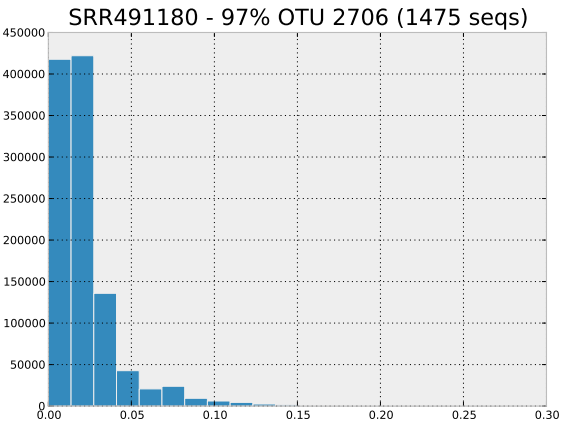
In order to understand the effect of UCLUST clustering at different similarity thresholds, we considered sample SRR491180 clustered at 97% and 85% of similarity and selected the first most abundant 7 OTUs in each case. Then, we computed the distances among sequences inside each OTUs and represented them in histograms for visualization. Finally we assigned taxonomy to each sequence inside these clusters to establish how well sequences were assembled together into clusters.

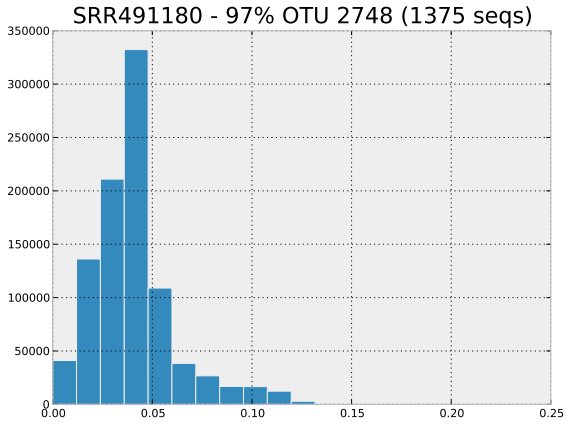
To compute the distance matrix, we first aligned our sequences through PyNAST, that is the python implementation of NAST (see 1.3.1). In order to align all sequences, without failures, we set the 'minimum sequence length to include in alignment' to 1 and the 'minimum percent sequence identity to include sequence in alignment' to 50%. Then we used mothur (see 1.3.3) with default settings to compute all distances. Finally, we represented the inside-OTU distances with histograms, as showed in the following figures.

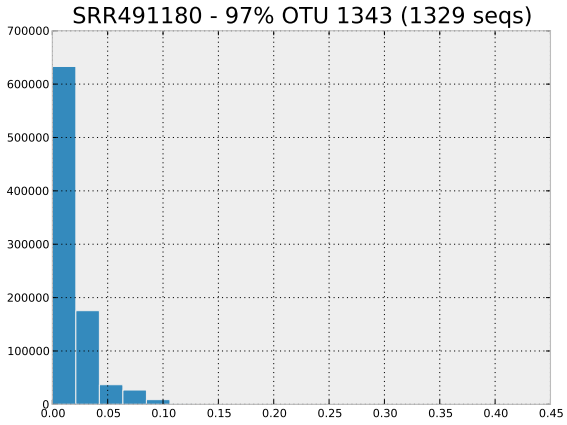

SRR491180 - 97% OTU 19 (3250 seqs)
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0.00
0.05
0.10
0.15
0.20

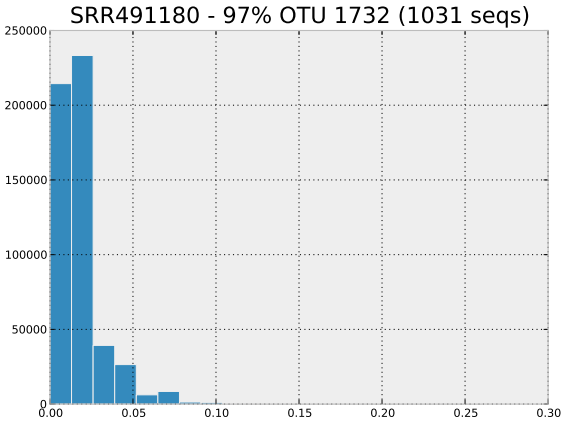

SRR491180 - 97% OTU 1805 (2002 seqs)
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
0.00
0.05
0.10
0.15
0.20
0.25

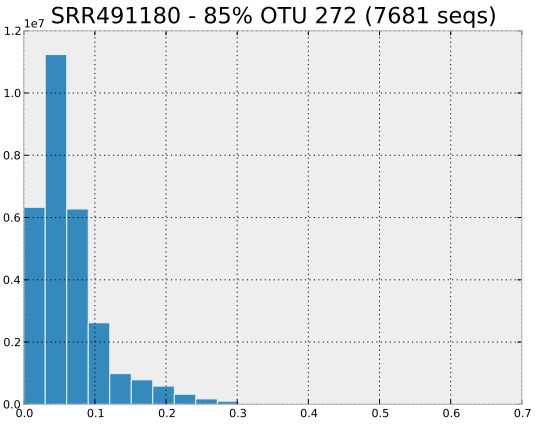

SRR491180 - 97% OTU 2710 (1614 seqs)
1000000
800000
600000
400000
200000
0
0.0
0.1
0.2
0.3
0.4
0.5

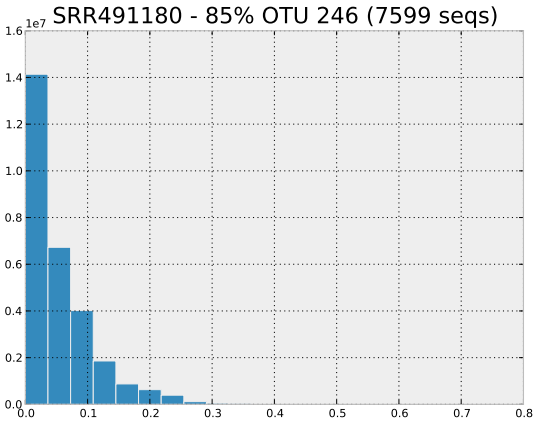

SRR491180 - 97% OTU 2706 (1475 seqs)
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30

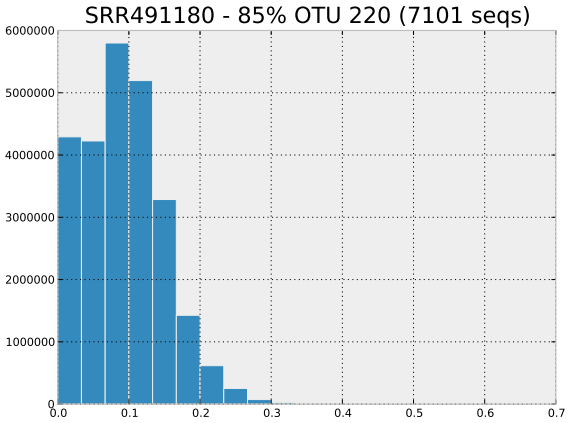

SRR491180 - 97% OTU 2748 (1375 seqs)
350000
300000
250000
200000
150000
100000
50000
0
0.00
0.05
0.10
0.15
0.20
0.25

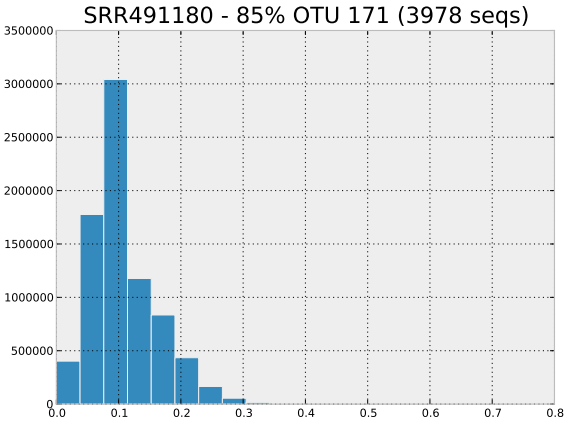

SRR491180 - 97% OTU 1343 (1329 seqs)
700000
600000
500000
400000
300000
200000
100000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

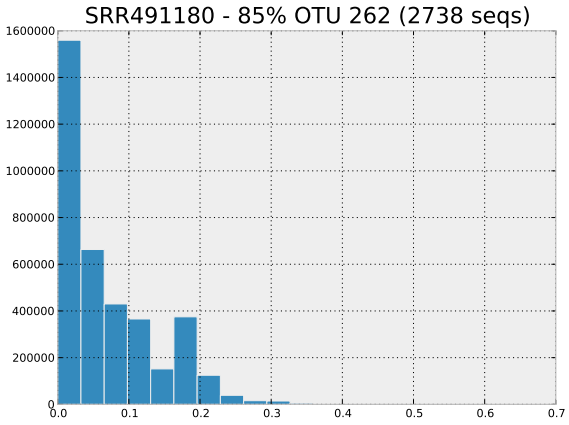

SRR491180 - 97% OTU 1732 (1031 seqs)
250000
200000
150000
100000
50000
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30

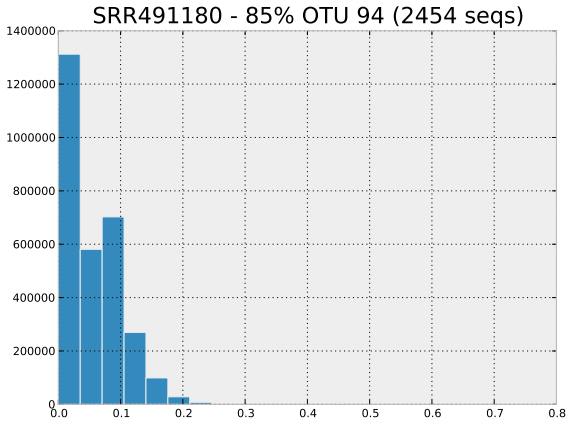

SRR491180 - 85% OTU 272 (7681 seqs)
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

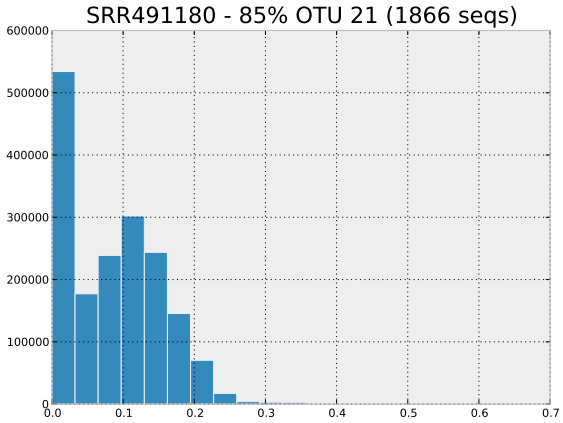

SRR491180 - 85% OTU 246 (7599 seqs)
1e7
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


SRR491180 - 85% OTU 220 (7101 seqs)
6000000
5000000
4000000
3000000
2000000
1000000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


SRR491180 - 85% OTU 171 (3978 seqs)
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


SRR491180 - 85% OTU 262 (2738 seqs)
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


SRR491180 - 85% OTU 94 (2454 seqs)
1400000
1200000
1000000
800000
600000
400000
200000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


SRR491180 - 85% OTU 21 (1866 seqs)
600000
500000
400000
300000
200000
100000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

From the previous figures we can see how well UCLUST builds its clusters even if it does not compare all vs all sequences and we can already underline the differences between clustering at high similarity thresholds (in this case 97%) or at low similarity thresholds (85%).
At 97% of similarity, in fact, we can see how, even in these very plentiful clusters, the maximum distance inside each cluster is approximately always under 0.1, in a scale from 0 to 1. Distance histograms show tails on the right, as expected, since distances can be thought as sums of squares of independent normal distributed variables, which will follow a chi-square distribution, that in fact exhibits a tail on the right. Furthermore, likely, histograms do not contain secondary peaks, which could have been a sign of multiple clusters inside the considered one.
At 85%, instead, histograms are more spread and maximum distances exceed 0.1 remaining however always approximately under 0.3. Also here, the shape of the distances distribution is, as expected, with a tail towards high distances and without secondary peaks.
These observations let us already note how clustering at 85% similarity generates bigger and more spread clusters, which will contain more variability than those obtained with the 97% similarity.

As we mentioned before, our second test on UCLUST consisted in computing the inside-OTU taxonomy. For this purpose, we exploited the RDP Classifier (see section 1.3.4) with a bootstrap cutoff of 50%. We report in the following tables the results of the same clusters used in the distances computation. For each cluster we report the OTU identification number, its abundance, that is the number of sequences collected in it, the taxonomic assignment of its representative sequence, the taxonomic assignments present among its elements at each taxonomic level (from B to G), and, for each of these levels, how many sequences of the cluster were classified with that specific assignment.

OTU 19 - 3250 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Psychrobacter

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	3250	Proteobacteria	3250	Gammaproteobacteria	3250	Pseudomonadales	3250	Moraxellaceae	3249	Psychrobacter	3249

OTU 1805 - 2002 seqs - Representative sequence: Bacteria

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	2002	Firmicutes	1016	Clostridia	901	Desulfovibrionales	16	Desulfohalobiaceae	16		
		Proteobacteria	34	Deltaproteobacteria	21	Clostridiales	667				

OTU 2710 - 1614 seqs - Representative sequence: Bacteria; Firmicutes; Bacilli; Lactobacillales; Carnobacteriaceae; Carnobacterium

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1611	Firmicutes	1611	Bacilli	1611	Lactobacillales	1608	Carnobacteriaceae	1607	Carnobacterium	1605
Unassignable	3									Desemzia	1

OTU 2706 - 1475 seqs - Representative sequence: Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Fastidiosipila

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1475	Firmicutes	1457	Clostridia	1451	Clostridiales	1450	Ruminococcaceae	1431	Fastidiosipila	773

OTU 2748 - 1375 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Buttiauxella

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1375	Proteobacteria	1375	Gammaproteobacteria	1375	Enterobacteriales	1372	Enterobacteriaceae	1372	Buttiauxella	1339

OTU 1343 - 1329 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1327	Proteobacteria	1323	Gammaproteobacteria	1323	Enterobacteriales	1315	Enterobacteriaceae	1315	Yersinia	885
Unassignable	2	Actinobacteria	3	Actinobacteria	3	Actinomycetales	3			Serratia	354

OTU 1732 - 1031 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Psychrobacter

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1031	Proteobacteria	1031	Gammaproteobacteria	1031	Pseudomonadales	1028	Moraxellaceae	1028	Psychrobacter	1027

Table 5.2: Taxonomy of the firsts most abundant 7 OTUs of SRR114980 at 97% similarity.

OTU 272 - 7681 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	7656	Proteobacteria	7548	Gammaproteobacteria	7489	Enterobacteriales	6926	Enterobacteriaceae	6926	Vibrio	2
Unassignable	25	Actinobacteria	25	Actinobacteria	25	Vibrionales	2	Vibrionaceae	2	Kluyvera	2
						Actinomycetales	23	Mycobacteriaceae	12	Erwinia	3
										Mycobacterium	12
										Citrobacter	740
										Trabulsiella	37
										Raoultella	17
										Salmonella	22
										Hafnia	9
										Escherichia Shigella	11
										Pantoea	1
										Xenorhabdus	391
										Buttiauxella	4386
										Enterobacter	58

OTU 246 - 7599 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Psychrobacter

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	7568	Firmicutes	1	Betaproteobacteria	1	Enterobacteriales	3	Enterobacteriaceae	3	Bartonella	15
Unassignable	31	Proteobacteria	7558	Clostridia	1	Rhizobiales	15	Bartonellaceae	15	Psychrobacter	7190
				Alphaproteobacteria	26	Clostridiales	1	Moraxellaceae	7249		
				Gammaproteobacteria	7481	Oceanospirillales	1				
						Pseudomonadales	7253				

OTU 220 - 7101 seqs - Representative sequence: Bacteria

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	7096	Firmicutes	2498	Clostridia	2241	Synergistales	86	Ruminococcaceae	140	Sedimentibacter	4
Unassignable	5	Proteobacteria	462	Synergistia	86	Desulfovibrionales	182	Synergistaceae	86	Desulfonatrono-spira	1
		Synergistetes	86	Deltaproteobacteria	279	Clostridiales	1652	Incertae Sedis XI	8	Pyramidobacter	7
								Desulfohalobiaceae	167	Dethiosulfovibrio	1
								Incertae Sedis XIII	5	Tindallia	2
								Clostridiaceae	5	Ethanoligenens	33
										Fastidiosipila	39
										Acetanaero-bacterium	5
										Desulfohalobium	1
										Anaerovorax	5

Table 5.3: Taxonomy of the first most abundant 7 OTUs of SRR114980 at 85% similarity (first part).

OTU 171 - 3978 seqs - Representative sequence: Bacteria; Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; Butyrivibrio

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	3977	Firmicutes	3946	Clostridia	3943	Clostridiales	3941	Incertae Sedis XIV	18	Sporacetigenium	8
Unassignable	1	Actinobacteria	5	Actinobacteria	5	Actinomycetales	5	Ruminococcaceae	3	Blautia	18
								Peptococcaceae	1	Syntrophococcus	63
								Lachnospiraceae	3818	Lachnobacterium	98
								Incertae Sedis XI	6	Moryella	20
								Clostridiaceae	3	Sporotomaculum	1
								Peptostreptococcaceae	10	Parasporobacterium	5
										Coprococcus	574
										Butyrivibrio	754
										Anaerobacter	3
										Shuttleworthia	3
										Pseudobutyrivibrio	30
										Gallicola	6
										Dorea	1
										Anaerosporobacter	56
										Roseburia	20

OTU 262 - 2738 seqs - Representative sequence: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	2734	Proteobacteria	2731	Gammaproteobacteria	2726	Enterobacteriales	2562	Enterobacteriaceae	2562	Rahnella	6
Unassignable	4									Raoultella	2
										Citrobacter	30
										Serratia	793
										Escherichia/ Shigella	1
										Yersinia	1248
										Xenorhabdus	2
										Buttiauxella	44

OTU 94 - 2454 seqs - Representative sequence: Bacteria; Firmicutes; Bacilli; Lactobacillales; Enterococcaceae

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	2452	Firmicutes	2451	Bacilli	2450	Lactobacillales	2330	Carnobacteriaceae	2020	Carnobacterium	1922
Unassignable	2					Bacillales	6	Bacillaceae	2	Enterococcus	230
								Enterococcaceae	247	Desemzia	2
										Trichococcus	93

Table 5.4: Taxonomy of the first most abundant 7 OTUs of SRR114980 at 85% similarity (second part).

OTU 21 - 1866 seqs - Representative sequence: Bacteria; Firmicutes; Clostridia; Clostridiales; Ruminococcaceae; Oscillibacter

B	n.seqs B	C	n.seqs C	D	n.seqs D	E	n.seqs E	F	n.seqs F	G	n.seqs G
Bacteria	1845	Firmicutes	1687	Clostridia	1678	Synergistales	2	Ruminococcaceae	1538	Acetivibrio	8
Unassignable	21	Proteobacteria	56	Synergistia	2	Nautiliales	5	Synergistaceae	2	Desulfonauticus	5
		Synergistetes	2	Epsilonproteobacteria	5	Desulfovibrionales	9	Lachnospiraceae	30	Clostridium	7
				Deltaproteobacteria	10	Clostridiales	1669	Clostridiaceae	8	Jonquetella	1
								Nautiliaceae	5	Robinsoniella	1
								Desulfohalobiaceae	8	Fastidiosipila	607
										Nautilia	5
										Oscillibacter	160

Table 5.5: Taxonomy of the first most abundant 7 OTUs of SRR114980 at 85% similarity (third part).

Also taxonomy results show how different similarity thresholds used in the clustering procedure influence the output clusters structure.
At 97% similarity, almost all elements in the same cluster are classified with the same taxonomic assignment by the RDP Classifier. Furthermore, we can test the goodness of the cluster seed choice comparing the representative sequence taxonomy to that of the other sequences inside the cluster. Except for cluster 1805, in which it is clear that the seed taxonomy was not well assigned (note that however also in this cluster distances are $\lesssim 0.1$, so the cluster is well built) and consequently also the taxonomy inside the cluster does not show good results, for the other clusters, the representative sequence is able to attract to itself only elements of the same genus and moreover this happens in these very abundant clusters, so we can suppose that the same behavior will characterize also smaller clusters.
At 85% similarity, we observe that clusters are more spread, as already resulted in distance histograms, and even if seed sequences are classified up to the most refined level (level G), we can assert that sequences in the same cluster agree in taxonomic assignments just up to level F. Also here we have an exception for OTU 220, in which the representative sequence is not well classified as just level B taxonomy was assigned, that is we just know we are dealing with a bacteria, and consequently also its sequences agree in the classification just at level B. Also in this OTU, however, distances are not higher than in other 85% clusters, so we can assert that the clustering was well done, even if the RDP Classifier did not find good results.

5.4 Preston plots and fits

After clustering the samples data with UCLUST (through QIIME) at different similarity thresholds, as explained before, we computed the species abundances. Here, as we already mentioned, a 'species' corresponds to an OTU, thus its abundance will be exactly the number of sequences which have fallen into that cluster. Then we represented the relative abundance distribution by building the histogram of the species abundances with the x-axes in log_2 scale, thus obtaining a Preston plot (see 4.2).
Finally we fitted the histogram using the gamma-like function

$$f(x) = a \cdot b^x \cdot x^c, \tag{5.2}$$

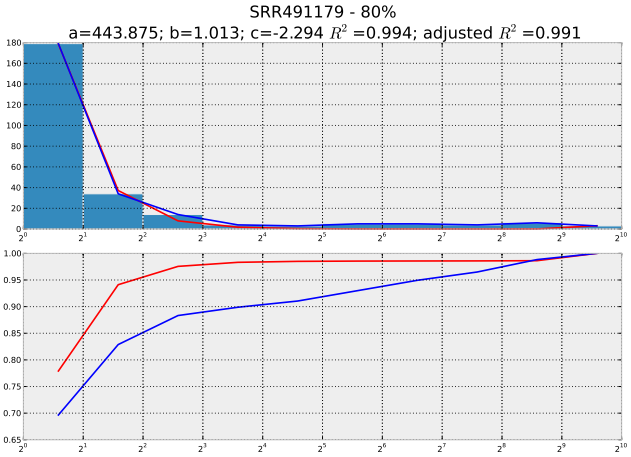
that corresponds to the gamma distribution 4.10 found in the continuous model 4.4, that we report here:

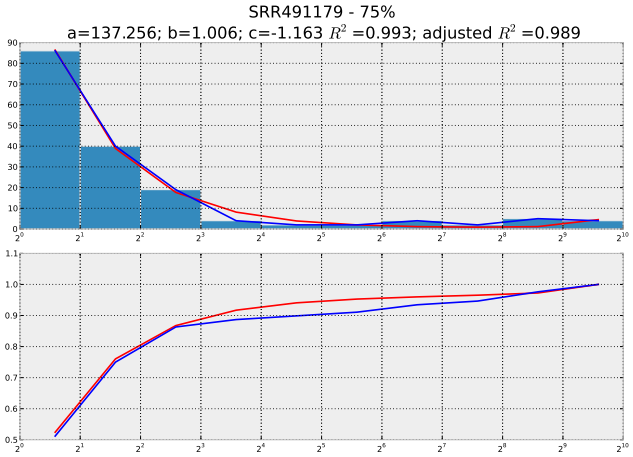
$$P_{RSA} = \frac{(D\tau)^{-b/D}}{\Gamma(b/D)} x^{\frac{b}{D}-1} e^{-\frac{x}{D\tau}}.$$

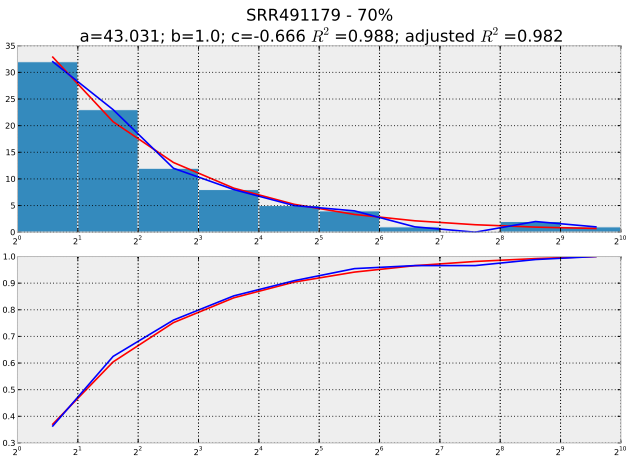
This gamma distribution itself can be considered as the continuous form of the negative binomial distribution 4.31 deduced by Volkov et al., that is

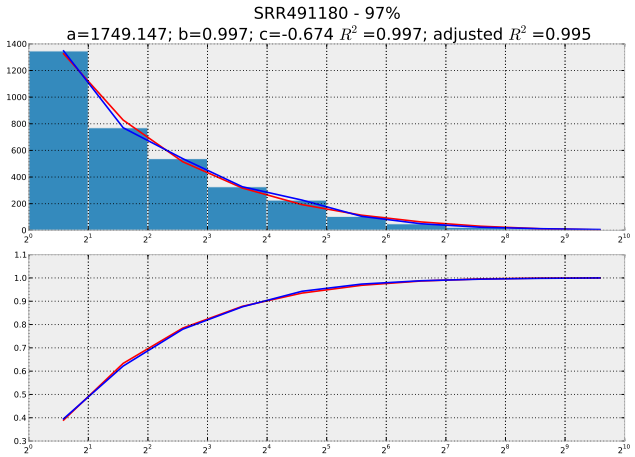
$$P_{n,k} = \frac{(1-x_k)^{\Upsilon_k}}{\Gamma(\Upsilon_k)} \frac{x_k^n}{n!} \Gamma(n+\Upsilon_k).$$

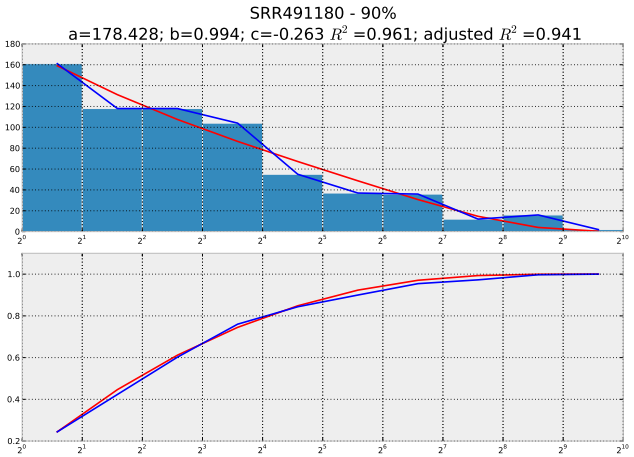
Thus, for simplicity we chose to fit our data with the simple function 5.2, that we can consider to be a distribution when its parameters satisfy the conditions $c+1>0 \rightarrow c>-1$ and $ln(1/b)>0 \rightarrow b<1$.
In particular we are interested in the gamma shape parameter $c+1$ that, as we remarked in section 4.4, is related to the immigration parameter Υ of the discrete model, or to the b/D parameter of the continuous model, which corresponds to an immigration term divided by a term that describes the fluctuations due to demographic stochasticity. Thus, the parameter c can help us understanding which is the dynamics under our system and what type of interactions are present among its constituents. Furthermore, following the example reported in section 4.5, we can perceive which type of community we are observing through our clustering procedure.
So, let us report our results, taking a particular attention to the shape of the fitted function, that is to the parameter c. In the following figures we show for each sample at each similarity threshold used, a first figure with the histogram in log_2 x-scale, the histogram points plot (blue line) and the fit curve (red line), and a second figure representing the cumulative function of the histogram points plot (blue) and that of the fit (red).

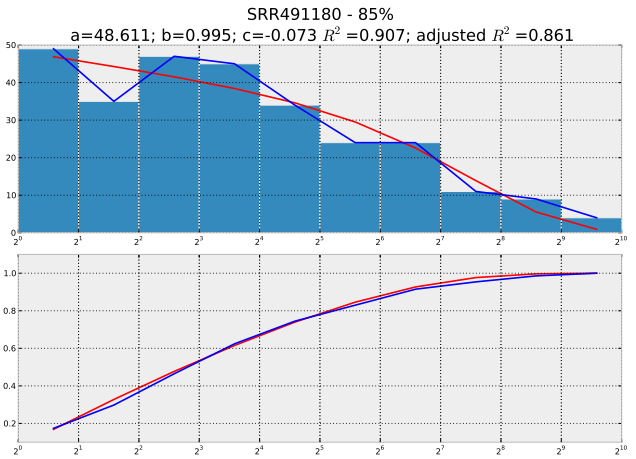

SRR491179 - 80%
a=443.875; b=1.013; c=-2.294 R^2 =0.994; adjusted R^2 =0.991

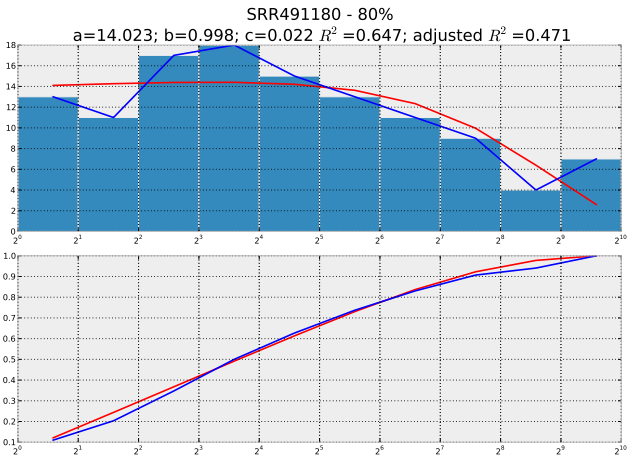

SRR491179 - 75%
a=137.256; b=1.006; c=-1.163 R^2 =0.993; adjusted R^2 =0.989

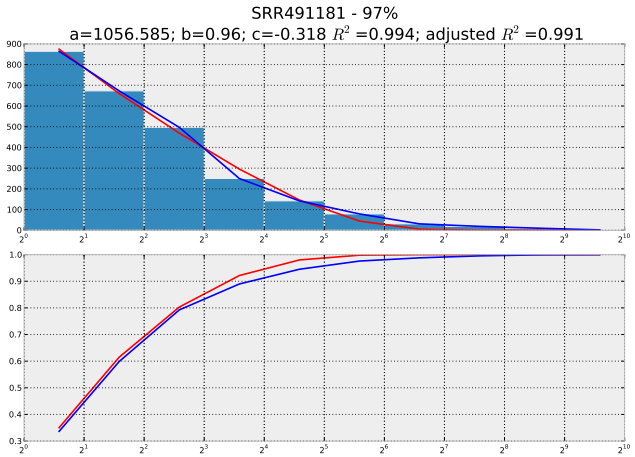

SRR491179 - 70%
a=43.031; b=1.0; c=-0.666 R^2 =0.988; adjusted R^2 =0.982

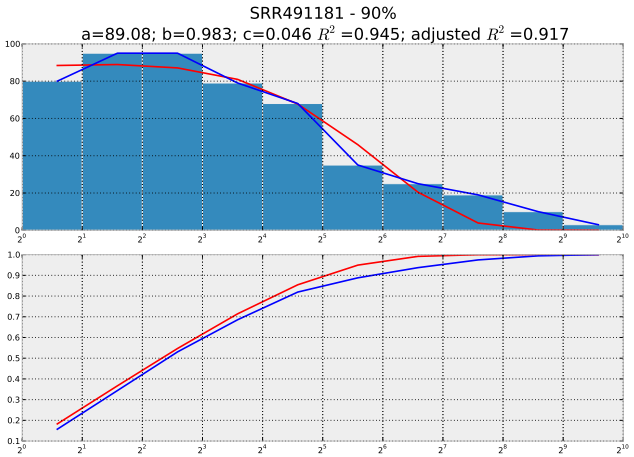

SRR491180 - 97%
a=1749.147; b=0.997; c=-0.674 R^2 =0.997; adjusted R^2 =0.995

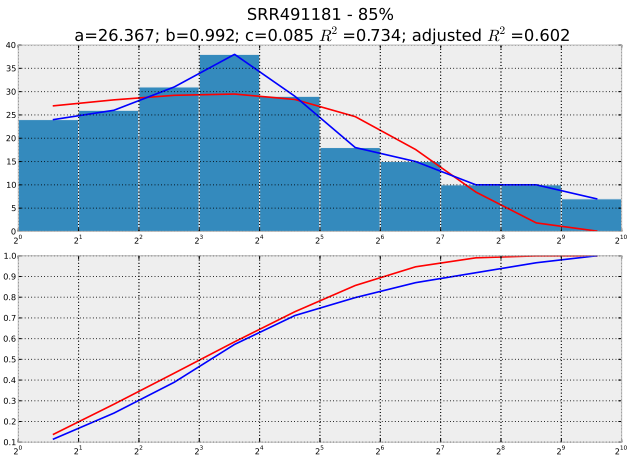

SRR491180 - 90%
a=178.428; b=0.994; c=-0.263 R^2 =0.961; adjusted R^2 =0.941

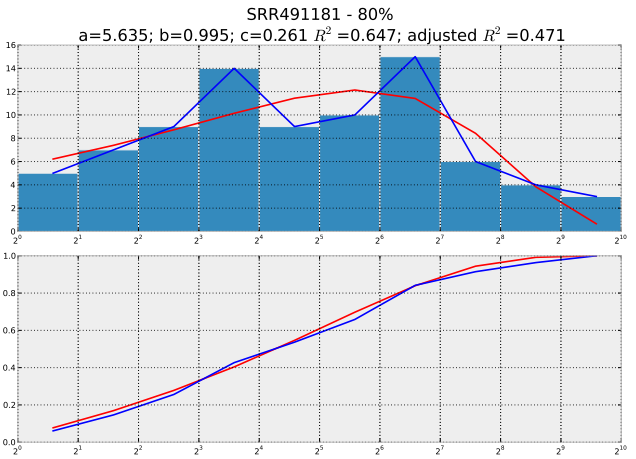

SRR491180 - 85%
a=48.611; b=0.995; c=-0.073 R^2 =0.907; adjusted R^2 =0.861

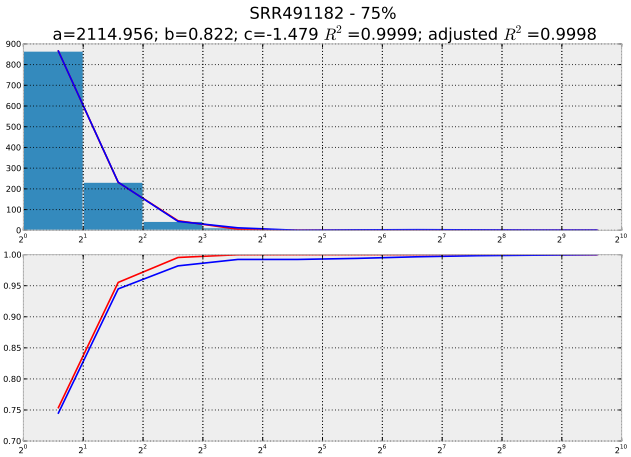

SRR491180 - 80%
a=14.023; b=0.998; c=0.022 R^2 =0.647; adjusted R^2 =0.471

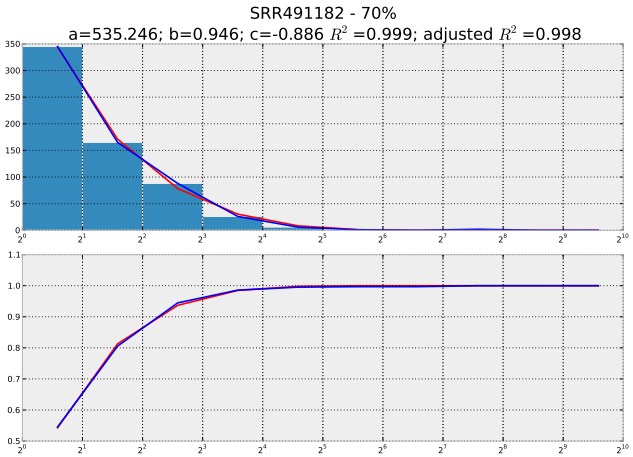

SRR491181 - 97%
a=1056.585; b=0.96; c=-0.318 R^2 =0.994; adjusted R^2 =0.991

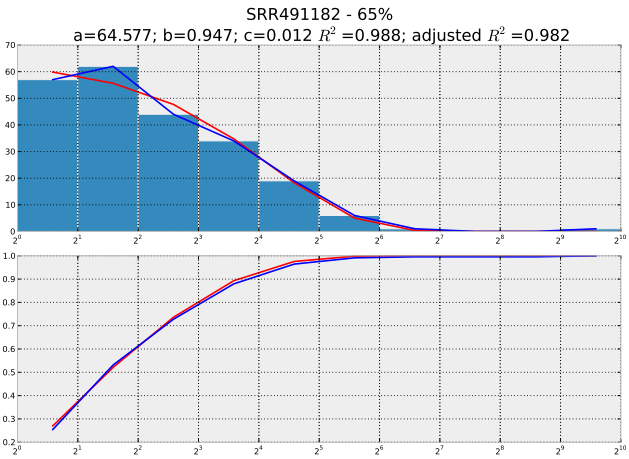

SRR491181 - 90%
a=89.08; b=0.983; c=0.046 R^2 =0.945; adjusted R^2 =0.917

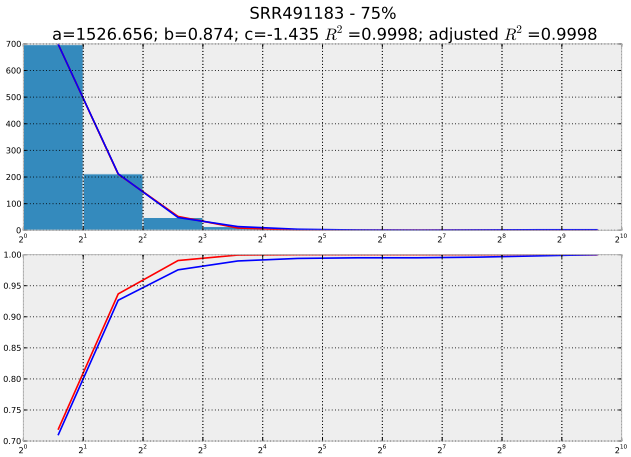

SRR491181 - 85%
a=26.367; b=0.992; c=0.085 R^2 =0.734; adjusted R^2 =0.602

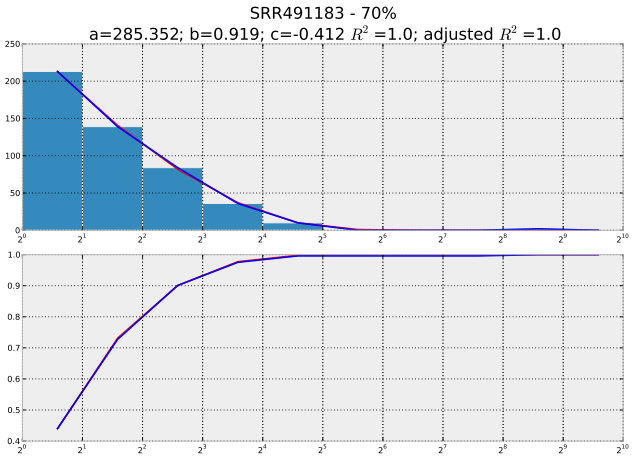

SRR491181 - 80%
a=5.635; b=0.995; c=0.261 R^2 =0.647; adjusted R^2 =0.471

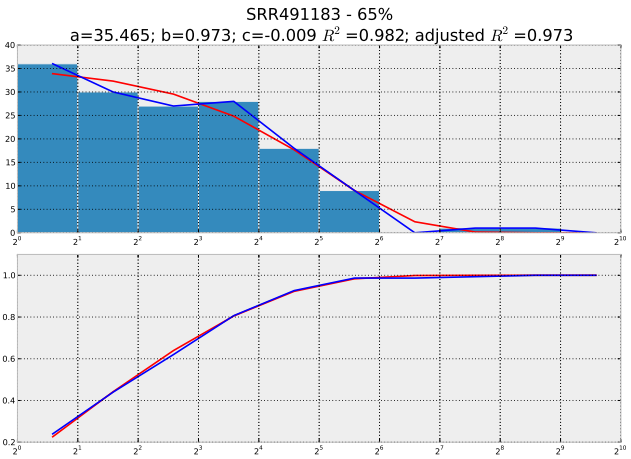

SRR491182 - 75%
a=2114.956; b=0.822; c=-1.479 R^2 =0.9999; adjusted R^2 =0.9998


SRR491182 - 70%
a=535.246; b=0.946; c=-0.886 R^2 =0.999; adjusted R^2 =0.998


SRR491182 - 65%
a=64.577; b=0.947; c=0.012 R^2 =0.988; adjusted R^2 =0.982


SRR491183 - 75%
a=1526.656; b=0.874; c=-1.435 R^2 =0.9998; adjusted R^2 =0.9998


SRR491183 - 70%
a=285.352; b=0.919; c=-0.412 R^2 =1.0; adjusted R^2 =1.0


SRR491183 - 65%
a=35.465; b=0.973; c=-0.009 R^2 =0.982; adjusted R^2 =0.973

Conclusions from results Looking at Preston plots results, we can note how well the theoretically predicted gamma distribution fits Jeraldo's gut microbiota sequencing data. We can note how, as the similarity threshold decreases, our Preston plots show a transition from a logserie-like to a lognormal-like curve.
This result of course reminds us that of the coral reefs (see section 4.5), in which this passage was explained in terms of local communities and metacommunities. As for the coral reefs, here we have an 'immigration' parameter c that increases as the similarity threshold decreases, as showed in figure 5.1. Note that in some cases we have $c < -1$, thus our function would not be exactly a distribution. Anyway we can give an interpretation considering these as limit cases with $c \sim -1$.

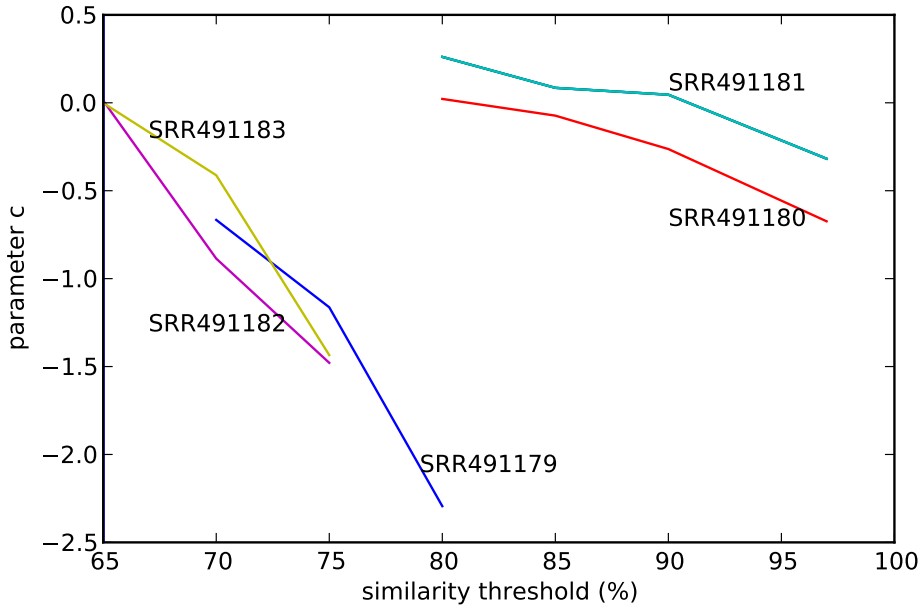


Figure 5.1: Trend of the parameter c for different samples. Note how c always decreases with the increase of the similarity threshold and how data from the same animal species produce similar c.

What we can assert is that when we cluster our data with a high similarity threshold, we are sampling many small OTUs (as already observed in distance histograms and in taxonomic assignments inside OTUs) which have a low immigration rate, that corresponds to having many small isolated local communities, separated from each other. In such a situation, the RSA resembles the Fisher logseries x^n/n (eq.4.1), since the gamma shape parameter $(c + 1)$ is very close to 0

($c \sim -1$).
With the decreasing of the similarity threshold, we sample fewer but bigger OTUs, as in our previous results, and the immigration rate increases. This is because, again referring to the coral reefs example, sampling bigger OTUs corresponds to assembling together multiple local communities into metacommunities. Thus, the RSA distribution of course acquires an interior mode, becoming more lognormal-like and the immigration parameter increases, since the many local communities assembled together will interact between each others and there will be immigration from each local community towards the others. In confirmation of this we can observe that with the decreasing of the similarity threshold, the gamma shape parameter c increases, as observed for the coral reefs.
Furthermore, from figure 5.1, we can also see how well different animal species gut microbiota result separated in the parameter c plot: SRR491180 and SRR491181 both refer to cattle rumen gut microbiota, SRR491182 and SRR491183 both refer to swine gut microbiota, while SRR491179 refers to chicken cecum microbiota.
To conclude, we can assert that the sampled microbiota community can be considered as an ensemble of isolated local communities, if we compute a refined analysis, that is if we look really close, while it shows a metacommunity behavior if we look at it more roughly, as also confirmed by our previous tests on inside-OTU distances distribution and inside-OTU taxonomic assignment. So, from our analysis we can extract informations on the structure and the dynamics of gut microbiota populations. Furthermore, we can also differentiate the species of origin of the gut microbiota and we can hope to apply these results in the understanding of the changing of gut microbiota populations in different conditions and pathologies.

Conclusions

In this work we analyzed next-generation sequencing data of 5 animals from [35], acquired with 454 Life Sciences pyrosequencing.
We compared the provided sequences through the algorithm UCLUST [2] and we thus generated OTUs, which correspond to bacteria species at particular taxonomic levels. We obtained clusters with different structures by using different threshold similarities. For high similarity we obtained many small clusters, which show a behavior similar to the coral-reef communities, when considered as isolated local communities, with a Relative Species Abundance (RSA) distribution well described by a log-serie like curve, as first proposed by Fisher [30]. For low similarity thresholds, instead, we could see the trend of the relative species abundance for fewer but bigger OTUs. In this case, the Preston plot acquired an interior mode and became lognormal-like.
This passage from a log-serie to a lognormal curve had been observed previously in many ecological systems and has been mathematically described with a simple dynamical model based on a master equation by Volkov et al. [49].
Such as in the work of Volkov et al. for the coral-reef community, we were able to extract informations on a shape parameter, which is related to some kind of interaction among species (OTUs) and which, for the coral reef, was considered as an immigration parameter. As expected, we obtained a shape parameter which increases with the decreasing of the similarity threshold, that is shifting from a refined view of isolated local communities (small clusters) to a rougher view of metacommunities, each composed of different local communities.
We proved that this model also works well for our sequencing data, showing how the gut microbiota community can be studied as an ecological system. This can help us understanding its biodiversity and the phenomenon of dysbiosis, which affects patients with metabolic diseases such as obesity or type 2 diabetes.
The technique of applying ecological models to sequencing data seems to have a great future. Further works will extend this ecological description to genes sequences in which species, families, etc. are characterized by the percentage of bases in common, even if they do not refer to microorganisms as for gut microbiota sequencing. In particular, a genetic ecosystem can be that of the transposable

elements (TEs) [47], that is very important in genetic and biodiversity regulation. These are more or less long bases sequences which raid DNA and whose invasion and replication mechanisms inside DNA itself constitute a complex and not entirely clear issue. Given that TEs have a considerable impact on the biology of their host species, we need to better understand whether their dynamics reflects some form of organization or is primarily driven by stochastic processes, and this can be of course another application of all the procedures, algorithms and models explained in this work.

Appendix A

Stochastic processes - Fluctuations

Let us consider a random variable x defined on $\mathbb{R}$, with probability density ρ, and let us consider the generation of N of these events: $x_1, ..., x_N$. We want to compute the mean number N_A of events that fall inside an ensemble $A \subset \mathbb{R}$ and in particular we are interested in its fluctuation $\frac{\sigma_A}{N_A}$. For this purpose we consider the random variable $y = \sum_{i=1}^{N} \chi_A(x_i)$, whose value corresponds exactly to the number of events that fall inside A (y is the number of times that x falls inside A), and where χ_A is the characteristic function of A: $\chi_A = \begin{cases} 0, & \text{if } x \notin A \\ 1, & \text{if } x \in A \end{cases}$.
In general, if $y = g(x)$ is a function of a random variable x with density function $\rho(x)$, the density function of y is given by $\hat{\rho}(y) = \int_{-\infty}^{\infty} \rho(x)\delta(y - g(x))dx$. Consequently the probability density function of y is

$$\hat{\rho}(y) = \int_{-\infty}^{\infty} \delta\left(y - \sum_{i=1}^{N} \chi_A(x_i)\right) \rho(x_1) \cdots \rho(x_N) dx_1 \cdots dx_N \tag{A.1}$$

Let us calculate the mean of y:

$$\begin{aligned} N_A =< y > &= \int y\hat{\rho}(y)dy \\ &= \int \rho(x_1)\cdots\rho(x_N)dx_1\cdots dx_N \int y\delta\left(y - \sum_{i=1}^{N}\chi_A(x_i)\right)dy \\ &= \sum_{i=1}^{N}\int \chi_A(x_i)\rho(x_1)\cdots\rho(x_N)dx_1\cdots dx_N \\ &= \sum_{i=1}^{N}\int \chi_A(x_i)\cdot\rho(x_i)dx_i \\ &= \sum_{i=1}^{N}\chi_A(x_i) \\ &= N\cdot\chi_A \end{aligned} \tag{A.2}$$

Where in the last steps we have applied the definition of χ_A: $\chi_A(x_i) \neq 0$ only if $x_i \in A$, and in this case we have $\int \chi_A(x_i)\rho(x_i)dx_i = \int_A \rho(x_i)dx_i = \begin{cases}\mu(x)\\ x\in A\end{cases} = \mu(A)$.
We now calculate the mean of y^2:

$$\begin{aligned} < y^2 > &= \int y^2\hat{\rho}(y)dy \\ &= \sum_{i,j=1}^{N}\int \chi_A(x_i)\chi_A(x_j)\rho(x_1)\cdots\rho(x_N)dx_1\cdots dx_N \end{aligned} \tag{A.3}$$

if $i = j$ we obtain a single integral on x_i: $\sum\int \rho(x_i)dx_i = N\mu(A)$;
if $i \neq j$ we will have a double integral: $\sum\int \rho(x_i)dx_i\rho(x_j)dx_j = N(N-1)\mu^2(A)$.
$< y^2 >$ is given by the sum of these two contributions: $< y^2 >= N(N-1)\mu^2(A) + N\mu(A)$.
Consequently, the variance will be

$$\sigma_A^2 =< y^2 > - < y >^2 = N\mu(A)(1-\mu(A)) \tag{A.4}$$

and it means that $\sigma_A \sim \sqrt{N}$.
We thus have that the fluctuation is given by

$$\frac{\sigma_A}{N_A} = \frac{1}{\sqrt{N}}\left[\frac{1-\mu(A)}{\mu(A)}\right]^{\frac{1}{2}} \sim \frac{1}{\sqrt{N}} \tag{A.5}$$

We can conclude that fluctuations go as $\frac{\sigma_A}{N_A} \sim \frac{1}{\sqrt{N}}$ (if $\mu(A) \sim 1$ we will have small fluctuations, while if A is small and thus $\mu(A) \ll 1$, fluctuations are likely to diverge) and that the stochastic process instead goes as $\sigma_A \sim \sqrt{N}$ [43].

Appendix B

DNA and RNA

DNA is the genetic material that organisms inherit from their parents. Encoded in the structure of DNA is the information that programs all the cell's activities, through the production of proteins. Each gene (stretch of DNA) along a DNA molecule directs synthesis of a type of RNA called messenger RNA (mRNA). The mRNA molecule interacts with the cells protein-synthesizing machinery to direct production of a polypeptide, which folds into all or part of a protein. We can summarize the flow of genetic information as $DNA \rightarrow RNA \rightarrow protein$. Actually, the system is much more complicated than this, involving many regulatory pattern, but for our description here, this simplification can be worthwhile.

B.1 DNA and RNA as molecules

Now, let us focus on the biochemical structure of DNA and RNA molecules.
DNA (DeoxyriboNucleic Acid) and RNA (RiboNucleic Acid) are a nucleic acid, and in particular they consist of long biopolymers made of simpler units called nucleotides. A nucleotide is composed of three parts: a nitrogen-containing (nitrogenous) base, a five-carbon sugar (a pentose), and one phosphate group (see fig.B.1).

To build a nucleotide, let us first consider the nitrogenous bases, described in fig.B.2. Then, let us add a sugar to the nitrogenous base. In DNA the sugar is deoxyribose, while in RNA it is ribose, as shown in fig.B.3. So far, we have built a nucleoside (nitrogenous base plus sugar). To complete the construction of a nucleotide, we attach a phosphate group to the 5' carbon of the sugar. The molecule is now a nucleoside monophosphate, better known as a nucleotide.
Now we can see how these nucleotides are linked together to build a polynu-

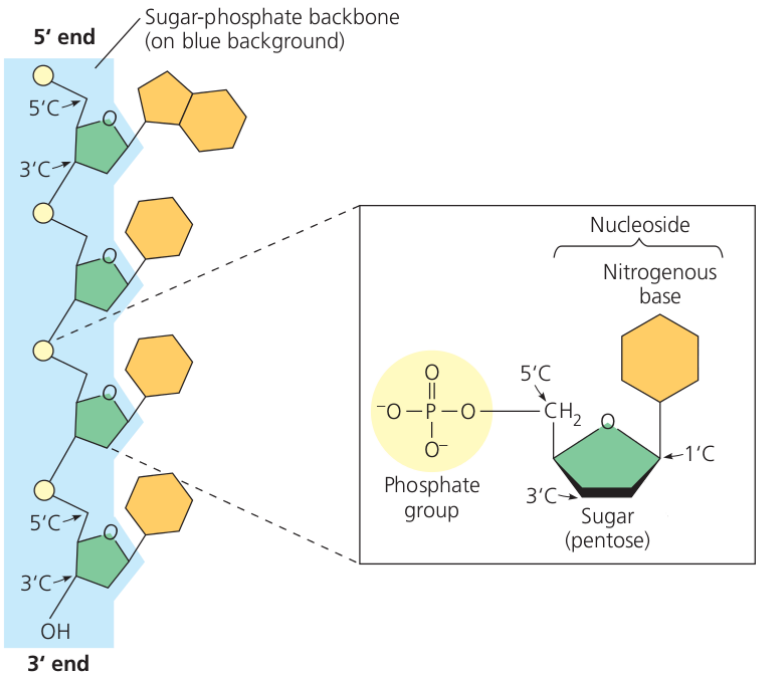


Figure B.1: (left) A polynucleotide has sugar-phosphate backbone with variable appendages, the nitrogenous bases. (right) A nucleotide monomer includes a nitrogenous base, a sugar, and a phosphate group. Figure from [45].

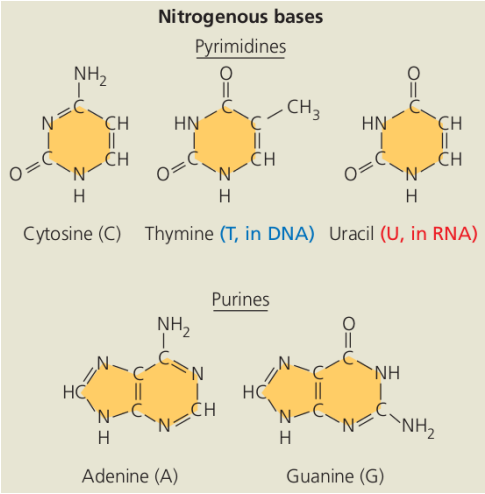


Figure B.2: Nitrogenous bases. Figure from [45].

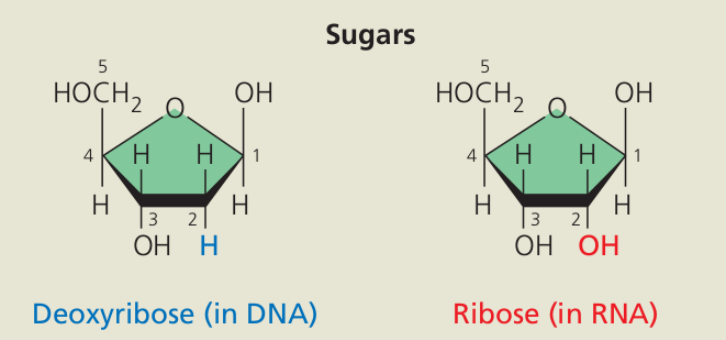


Figure B.3: Sugars. Figure from [45].

cleotide. Adjacent nucleotides are joined by a phosphodiester linkage, which consists of a phosphate group that links the sugars of two nucleotides. This bonding results in a backbone with a repeating pattern of sugar-phosphate units (see fig. B.1). The two free ends of the polymer are distinctly different from each other. One end has a phosphate attached to a 5' carbon, and the other end has a hydroxyl group on a 3' carbon; we refer to these as the 5' end and the 3' end, respectively. We can say that a polynucleotide has a built-in directionality along its sugar-phosphate backbone, from 5' to 3'.
RNA molecules usually exist as single polynucleotide chains like the one shown in fig.B.1. In contrast DNA molecules are double-stranded helices, consisting of two long biopolymers that spiral around an imaginary axis (fig.B.4).

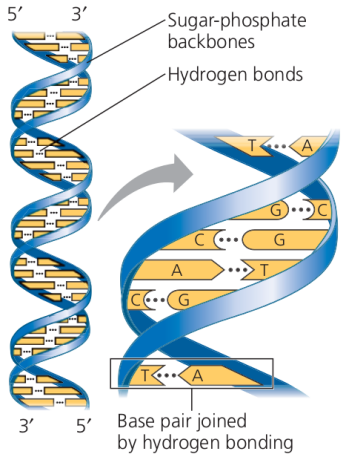


Figure B.4: DNA double elix. Figure from [45].

The two sugar-phosphate backbones run in opposite $5' \rightarrow 3'$ directions from each other; this arrangement is referred to as antiparallel. The sugar-phosphate backbones are on the outside of the helix, and the nitrogenous bases are paired in the interior of the helix. The two strands are held together by hydrogen bonds between the paired bases (see fig.B.4). Only certain bases in the double helix are compatible with each other. Adenine (A) always pairs with thymine (T), and guanine (G) always pairs with cytosine (C). Thus, the two strands of the double helix are complementary. Note that in RNA, adenine (A) pairs with uracil (U).

B.2 DNA replication

Let us now describe the process of DNA replication, that is exploited in many sequencing techniques.
The replication of a DNA molecule begins at particular sites called origins of replication, short stretches of DNA having a specific sequence of nucleotides. While many bacteria's chromosomes have just a single origin, eukaryotic chromosome may have hundreds or even a few thousand replication origins.
Proteins that initiate DNA replication recognize this origin sequence and attach to the DNA, separating the two strands and opening up a replication 'bubble'. Replication of DNA then proceeds in both directions until the entire molecule is copied. In eukaryotes, multiple replication bubbles form and eventually fuse, thus speeding up the copying of the very long DNA molecules.

At each end of a replication bubble is a replication fork, a Y-shaped region where the parental strands of DNA are being unwound. Several kinds of proteins participate in the unwinding (fig.B.5). Helicases are enzymes that untwist the double helix at the replication forks, separating the two parental strands and making them available as template strands. After the parental strands separate, single-strand binding proteins bind to the unpaired DNA strands, keeping them from re-pairing. The untwisting of the double helix causes tighter twisting and strain ahead of the replication fork. Topoisomerase helps relieve this strain by breaking, swiveling, and rejoining DNA strands.
The unwound sections of parental DNA strands are now available to serve as templates for the synthesis of new complementary DNA strands. However, the enzymes that synthesize DNA cannot initiate the synthesis of a polynucleotide; they can only add nucleotides to the end of an already existing chain that is base-paired with the template strand. The initial nucleotide chain that is produced during DNA synthesis is actually a short stretch of RNA, not DNA. This RNA chain is called a primer and is synthesized by the enzyme primase. Primase starts a complementary

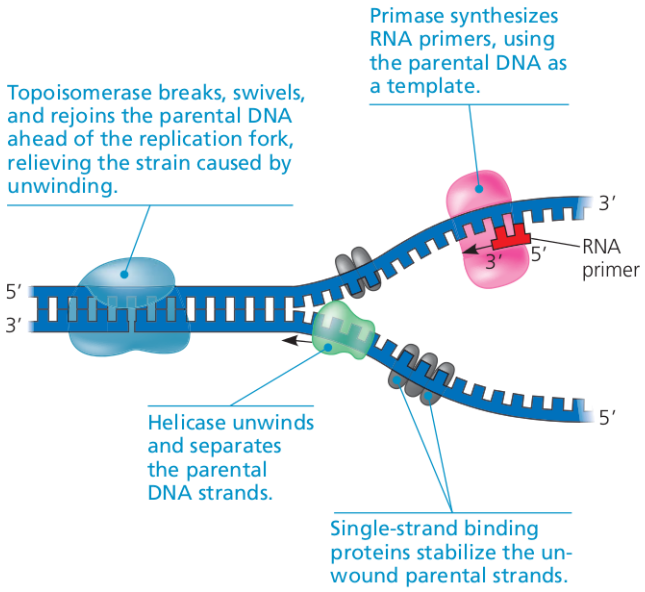


Figure B.5: Some of the proteins involved in the initiation of DNA replication. The same proteins function at both replication forks in a replication bubble. Figure from [45].

RNA chain from a single RNA nucleotide, adding RNA nucleotides one at a time, using the parental DNA strand as a template. The completed primer, generally 510 nucleotides long, is thus base-paired to the template strand. The new DNA strand will start from the 3' end of the RNA primer.

Enzymes called DNA polymerases catalyze the synthesis of new DNA by adding nucleotides to a preexisting chain. In E. coli, there are several different DNA polymerases, but two appear to play the major roles in DNA replication: DNA polymerase III and DNA polymerase I. The situation in eukaryotes is more complicated, with at least 11 different DNA polymerases discovered so far; however, the general principles are the same. Most DNA polymerases require a primer and a DNA template strand, along which complementary DNA nucleotides line up. Each nucleotide added to a growing DNA strand comes from a nucleoside triphosphate, which is a nucleoside (a sugar and a base) with three phosphate groups. The nucleoside triphosphates used for DNA synthesis are chemically reactive, partly because their triphosphate tails have an unstable cluster of negative charge. As each monomer joins the growing end of a DNA strand, two phosphate groups are lost as a molecule of pyrophosphate. Subsequent hydrolysis of the pyrophosphate to two molecules of inorganic phosphate is a coupled exergonic reaction that helps drive the polymerization reaction (fig.B.6).

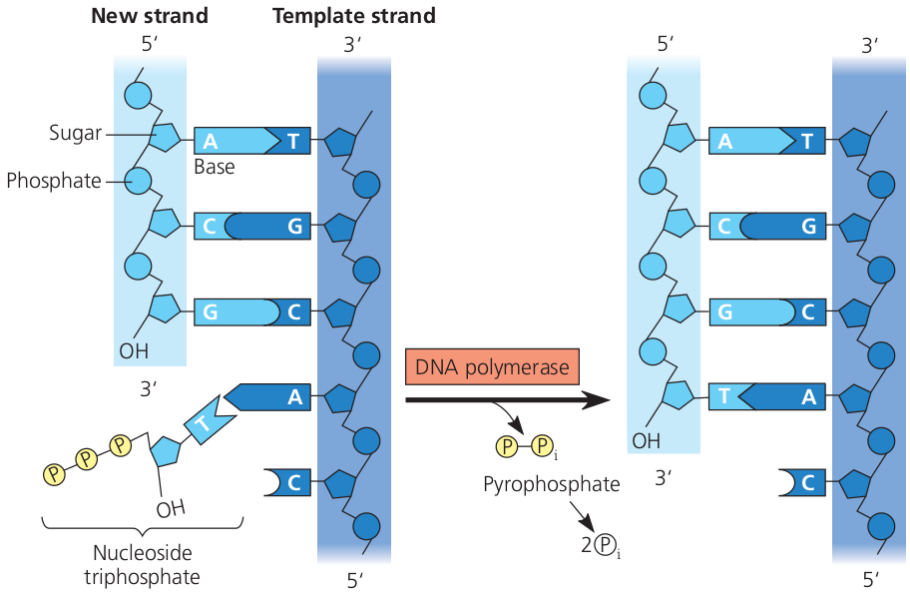


Figure B.6: Incorporation of a nucleotide into a DNA strand. Figure from [45].

Appendix C

16S ribosomal RNA and phylogenetic analysis

Evolution is a process whereby populations are altered over time and may split into separate branches, hybridize together, or terminate by extinction. The evolutionary branching process may be depicted as a phylogenetic tree, and the place of each of the various organisms on the tree is based on a hypothesis about the sequence in which evolutionary branching events occurred.
Phylogenetic analyses, that is the study of evolutionary relationships among groups of organisms (e.g. species, populations), which are discovered through molecular sequencing data and morphological data matrices, have become essential to research on the evolutionary tree of life.
Taxonomy, that is the classification, identification, and naming of organisms, is usually richly informed by phylogenetics. So, to finally understand why we do sequence 16S rRNA to obtain the classification of gut microbiota's bacteria populations, let us first explain how biologists classify living beings in pylogenetic trees and how ribosomal RNA can be exploited for this purpose.

C.1 The tree of life

Early taxonomists classified all known species into two kingdoms: plants and animals. Even with the discovery of the diverse microbial world, the two-kingdom system persisted: noting that bacteria had a rigid cell wall, taxonomists placed them in the plant kingdom. Eukaryotic unicellular organisms with chloroplasts were also considered plants. Fungi, too, were classified as plants, partly because most fungi, like most plants, are unable to move about (never mind the fact that fungi are not photosynthetic and have little in common structurally with

plants!). In the two-kingdom system, unicellular eukaryotes that move and ingest food (protozoans) were classified as animals. Those such as Euglena that move and are photosynthetic were claimed by both botanists and zoologists and showed up in both kingdoms. Taxonomic schemes with more than two kingdoms gained broad acceptance in the late 1960s, when many biologists recognized five kingdoms: Monera (prokaryotes), Protista (a diverse kingdom consisting mostly of unicellular organisms), Plantae, Fungi, and Animalia. This system highlighted the two fundamentally different types of cells, prokaryotic and eukaryotic, and set the prokaryotes apart from all eukaryotes by placing them in their own kingdom, Monera.
However, phylogenies based on genetic data soon began to reveal a problem with this system: some prokaryotes differ as much from each other as they do from eukaryotes. Such difficulties have led biologists to adopt a three-domain system. The three domains (Bacteria, Archaea, and Eukarya) are a taxonomic level higher than the kingdom level. The validity of these domains is supported by many studies, including a recent study that analyzed nearly 100 completely sequenced genomes.
The domain Bacteria contains most of the currently known prokaryotes, including the bacteria closely related to chloroplasts and mitochondria. The second domain, Archaea, consists of a diverse group of prokaryotic organisms that inhabit a wide variety of environments. Bacteria differ from archaea in many structural, biochemical, and physiological characteristics. The third domain, Eukarya, consists of all the organisms that have cells containing true nuclei.

Figure C.1 represents one possible phylogenetic tree for the three domains and the many lineages they encompass. The three-domain system highlights the fact that much of the history of life has been about single-celled organisms. The two prokaryotic domains consist entirely of single-celled organisms, and even in Eukarya, only the branches shown in red (plants, fungi, and animals) are dominated by multicellular organisms. Of the five kingdoms previously recognized by taxonomists, most biologists continue to recognize Plantae, Fungi, and Animalia, but not Monera and Protista. The kingdom Monera is obsolete because it would have members in two different domains. The kingdom Protista has also crumbled because it is polyphyletic: it includes members that are more closely related to plants, fungi, or animals than to other protists.
In the tree in fig.C.1, the first major split in the history of life occurred when bacteria diverged from other organisms. Thus, eukaryotes and archaea are more closely related to each other than either is to bacteria.
This reconstruction of the tree of life is based largely on sequence comparisons of rRNA genes, which code for the RNA components of ribosomes. Because ribosomes are fundamental to the workings of the cell, rRNA genes have evolved

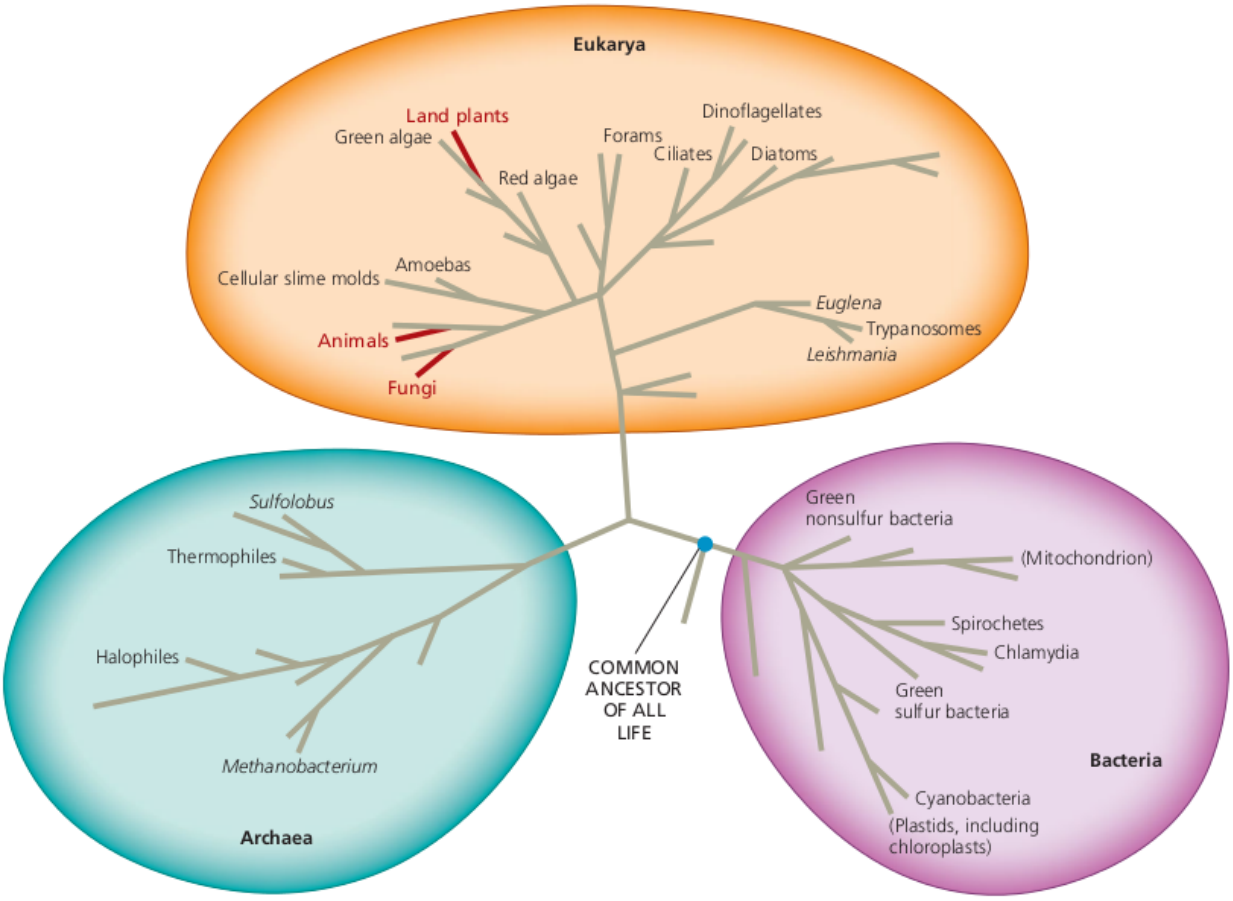


Figure C.1: The three domains of life. The phylogenetic tree shown here is based on rRNA gene sequences. Branch lengths are proportional to the amount of genetic change in each lineage. Figure from [45].

so slowly that homologies between distantly related organisms can still be detected, making these genes very useful for determining evolutionary relationships between deep branches in the history of life.

However, other genes reveal a different set of relationships. For example, researchers have found that many of the genes that influence metabolism in yeast (a unicellular eukaryote) are more similar to genes in the domain Bacteria than they are to genes in the domain Archaea, a finding that suggests that the eukaryotes may share a more recent common ancestor with bacteria than with archaea. Comparisons of complete genomes from the three domains show that there have been substantial movements of genes between organisms in the different domains. These took place through horizontal gene transfer, a process in which genes are

transferred from one genome to another through mechanisms such as exchange of transposable elements and plasmids, viral infection, and perhaps fusions of organisms. Recent research reinforces the view that horizontal gene transfer is important. For example, a 2008 analysis indicated that, on average, 80% of the genes in 181 prokaryotic genomes had moved between species at some point during the course of evolution. Because phylogenetic trees are based on the assumption that genes are passed vertically from one generation to the next, the occurrence of such horizontal transfer events helps to explain why trees built using different genes can give inconsistent results.

C.2 16S ribosomal RNA

The ribosome is a large and complex molecular machine, found within all living cells, that serves as the primary site of biological protein synthesis (translation). Ribosomes link amino acids together in the order specified by messenger RNA (mRNA) molecules. Ribosomes consist of two major subunits: the small ribosomal subunit reads the mRNA, while the large subunit joins amino acids to form a polypeptide chain. Each subunit is composed of one or more ribosomal RNA (rRNA) molecules and a variety of proteins.
Ribosomal RNA is suitable for phylogenetic studies since it is a component of all self-replicating systems, it is readily isolated and its sequence changes but slowly with time, permitting the detection of relatedness among very distant species. In particular, 16S ribosomal RNA (or 16S rRNA), which is a component of the 30S small subunit of prokaryotic ribosomes, is exploited for this purpose. 16S rRNA is 1.542kB (1542 nucleotides) in length and the genes coding for it, referred to as 16S rDNA, are indeed used in reconstructing phylogenies, thanks to the work of Carl Woese and George E. Fox [51].
Multiple sequences of 16S rRNA can exist within a single bacterium. The most common primer pair was devised by Weisburg et al. [50] and is currently referred to as 27F and 1492R; however, for some applications shorter amplicons may be necessary for example for 454 sequencing with Titanium chemistry (500-ish reads are ideal) the primer pair 27F-534R covering V1 to V3. Often 8F is used rather than 27F. Fig.C.2 shows these primers.
Type strains of 16S rRNA gene sequences for most bacteria and archaea are available on public databases such as NCBI. However, the quality of the sequences found on these databases are often not validated. Therefore, secondary databases which collect only 16S rRNA sequences are widely used. The most frequently used online databases are: EzTaxon-e, Ribosomal Database Project (RDP), SILVA and Greengenes.

Primer name	Sequence (5'-3')
8F	AGA GTT TGA TCC TGG CTC AG
U1492R	GGT TAC CTT GTT ACG ACT T
928F	TAA AAC TYA AAK GAA TTG ACG GG
336R	ACT GCT GCS YCC CGT AGG AGT CT
1100F	YAA CGA GCG CAA CCC
1100R	GGG TTG CGC TCG TTG
337F	GAC TCC TAC GGG AGG CWG CAG
907R	CCG TCA ATT CCT TTR AGT TT
785F	GGA TTA GAT ACC CTG GTA
805R	GAC TAC CAG GGT ATC TAA TC
533F	GTG CCA GCM GCC GCG GTA A
518R	GTA TTA CCG CGG CTG CTG G
27F	AGA GTT TGG ATC MTG GCT CAG
1492R	CGG TTA CCT TGT TAC GAC TT

Figure C.2: 16S ribosomal RNA primers. Figure from [16].

Bibliography

[1] Cd-hit user's guide, May 2009. URL http://weizhong-lab.ucsd.edu/cd-hit/.

[2] Uclust, extreme high-speed sequence clustering, alignment and database search, 2010. URL http://www.drive5.com/uclust.

[3] Dist.seqs, 2011. URL http://www.mothur.org/wiki/Dist.seqs.

[4] Python for bioinformatics. qiime (4) pynast, 2011. URL http://telliott99.blogspot.it/2011/02/qiime-4-pynast.html.

[5] Dna sequencing, 2013. URL http://en.wikipedia.org/wiki/DNA_sequencing.

[6] Maxamgilbert sequencing, 2013. URL http://en.wikipedia.org/wiki/Maxam-Gilbert_sequencing.

[7] Sequence read archive (sra), 2013. URL http://www.ncbi.nlm.nih.gov/sra.

[8] Illuminia, 2013. URL www.illuminia.com.

[9] Life technologies, 2013. URL http://www.lifetechnologies.com.

[10] Cluster, 2013. URL http://www.mothur.org/wiki/Cluster.

[11] Phred quality score, october 2013. URL http://en.wikipedia.org/wiki/Phred_quality_score.

[12] Qiime (quantitative insights into microbial ecology), 2013. URL http://qiime.org/.

[13] Assign taxonomy to each sequence, 2013. URL http://qiime.org/scripts/assign_taxonomy.html.

[14] Rdp, 2013. URL http://rdp.cme.msu.edu.

[15] Smithwaterman algorithm, 2013. URL http://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm.

[16] 16s ribosomal rna, november 2013. URL http://en.wikipedia.org/wiki/16S_ribosomal_RNA.

[17] 454 life sciences, 2013. URL http://en.wikipedia.org/wiki/454_Life_Sciences.

[18] Blast, 2013. URL http://en.wikipedia.org/wiki/BLAST.

[19] Single-linkage clustering, 2013. URL https://en.wikipedia.org/wiki/Single-linkage_clustering.

[20] G. Aprea, G. Gianese, V. Rosato, and V. Spedaletti. Il ruolo dell'ict nelle scienze omiche high-throughput. *EAI*, Speciale:74–82, 2013.

[21] S. Azaele, S. Pigolotti, J.R. Banavar, and A. Maritan. Dynamical evolution of ecosystems. *Nature*, 2006.

[22] S.C. Baker. Next-Generation Sequencing vs. Microarrays. Is it time to switch? *Genetic Engineering and Biotechnology News*, 2013.

[23] F. Bckhed, J.K. Manchester, C.F. Semenkovich, and J.I. Gordon. Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3):979–84, 2007.

[24] P.D. Cani, A.M. Neyrinck, F. Fava, C. Knauf, R.G. Burcelin, K.M. Tuohy, G.R. Gibson, and N.M. Delzenne. Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia*, 50:2374–2383, 2007.

[25] P.D. Cani, R. Bibiloni, C. Knauf, A.M. Neyrinck, and N.M. Delzenne. Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat dietinduced obesity and diabetes in mice. *Diabetes*, 57(June):1470–81, 2008.

[26] Marcus J Claesson, Orla O'Sullivan, Qiong Wang, Janne Nikkilä, Julian R Marchesi, Hauke Smidt, Willem M de Vos, R Paul Ross, and Paul W O'Toole. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PloS one*, 4(8):e6669, 2009.

[27] T.Z. DeSantis Jr, P. Hugenholtz, K. Keller, E.L. Brodie, N. Larsen, Y.M. Piceno, R. Phan, and G.L. Andersen. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, 34(2):W394–W399, 2006.

[28] R.C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 35(2):17921797, 2004.

[29] R.C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[30] R.A. Fisher, A.S. Corbet, and C.B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 1943.

[31] A. Gega and M.J. Kozal. New technology to detect low-level drug-resistant hiv variants. *Future Virology*, 6(1):17–26, 2011.

[32] E. Giampieri. *Stochastic models and dynamic measures for the characterization of bistable circuits in cellular biophysics*. PhD thesis, Facoltà di scienze matematiche, fisiche e naturali - University of Bologna, 2012.

[33] A. Gilles, E. Meglcz, N. Pech, S. Ferreira, T. Malausa, and J.F. Martin. Accuracy and quality assessment of 454 gs-flx titanium pyrosequencing. *BMC Genomics*, 12:245–255, 2011.

[34] S.P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton Univesity Press, 2001.

[35] P. Jeraldo, M. Sipos, N. Chia, J.M. Brulc, A.S. Dhillon, M.E. Konkel, C.L. Larson, et al. Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 109(25): 9692–8, 2012.

[36] N. Larsen, F.K. Vogensen, F.W.J. Van den Berg, D.S. Nielsen, A.S. Andreasen, B.K. Pedersen, W.A. Al-Soud, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS one*, 5(2): e9085, 2010.

[37] R.H. MacArthur and E.O. Wilson. *The Theory of Island Biogeography.* Princeton Univesity Press, 1967.

[38] Diana Marco. *Metagenomics. Current Innovations and Future Trends*. Caister Academic Press, 2011.

[39] A.J. Myers. The age of the ome: Genome, transcriptome and proteome data set collection and analysis. *Brain Research Bulletin*, 88:294–301, 2012.

[40] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[41] C. Saccone and G. Pesole. *Handbook of Comparative Genomics: Principles and Methodology.* Wiley, 2003.

[42] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[43] G. Turchetti. *Modelli Numerici della Fisica.* 2007.

[44] P.J. Turnbaugh, R.E. Ley, M. Mahowald, V. Magrini, E.R. Mardis, and J.I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31, 2006.

[45] L.A. Urry, M.L. Cain, S.A. Wasserman, and P.V. Minorsky. *Campbell biology - ninth edition*. Pearson Benjamin Cummings, 2011.

[46] N.G. Van Kampen. *Stochastic Processes In Physics And Chemistry.* Elsevier, 1981.

[47] S. Venner, C. Feschotte, and C. Bimont. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in genetics*, 25(7):317–23, 2009.

[48] J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers, and H.O. Smith. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[49] I. Volkov, J.R. Banavar, S.P. Hubbell, and A. Maritan. Patterns of relative species abundance in rainforests and coral reefs. *Nature*, 2007.

[50] W.G. Weisburg, S.M. Barns, D. Pelletier, and D.J. Lane. 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, 173(2): 697–703, 1991.

[51] C.R. Woese and G.E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, 1977.

[52] C. Zhang, M. Zhang, S. Wang, R. Han, Y. Cao, W. Hua, Y. Mao, et al. Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *The ISME journal*, 4(2):232–41, 2010.