Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea Magistrale in Fisica

Stochastic modeling of bacterial protein domains distribution to predict horizontal gene transfer

Relatore:
Prof. Gastone Castellani

Correlatore:
Prof. Daniel Remondini

Presentata da:
Marco Tamburi

Sessione I
Anno Accademico 2014/2015

Abstract

In questo elaborato, abbiamo tentato di modellizzare i processi che regolano la presenza dei domini proteici. I domini proteici studiati in questa tesi sono stati ottenuti dai genomi batterici disponibili nei data base pubblici (principalmente dal National Centre for Biotechnology Information: NCBI) tramite una procedura di simulazione computazionale. Ci siamo concentrati su organismi batterici in quanto in essi la presenza di geni trasmessi orizzontalmente, ossia che parte del materiale genetico non provenga dai genitori, è assodato che sia presente in una maggiore percentuale rispetto agli organismi più evoluti.

Il modello usato si basa sui processi stocastici di nascita e morte, con l'aggiunta di un parametro di migrazione, usato anche nella descrizione dell'abbondanza relativa delle specie in ambito delle biodiversità ecologiche.

Le relazioni tra i parametri, calcolati come migliori stime di una distribuzione binomiale negativa rinormalizzata e adattata agli istogrammi sperimentali, ci induce ad ipotizzare che le famiglie batteriche caratterizzate da un basso valore numerico del parametro di immigrazione abbiano contrastato questo deficit con un elevato valore del tasso di nascita. Al contrario, ipotizziamo che le famiglie con un tasso di nascita relativamente basso si siano adattate, e in conseguenza, mostrano un elevato valore del parametro di migrazione.

Inoltre riteniamo che il parametro di migrazione sia direttamente proporzionale alla quantità di trasferimento genico orizzontale effettuato dalla famiglia batterica.

Acknowledgments

Innanzitutto volevo ringraziare chi mi ha seguito in questo specifico elaborato, ossia il prof. Castellani e il prof. Remondini, i quali hanno indubbiamente trasmesso conoscenza e fatto entrare in un percorso di ricerca scientifica. Un ulteriore pensiero va alla dott.ssa Claudia, che è stata fondamentale per la parte iniziale del mio lavoro, e al gruppo del WUR, che sono stati sempre disponibili e cordiali.

Ringrazio i *Worsts* perché, benché negli ultimi anni siano diminuite fino a scomparire, le pause con voi erano un toccasana.

Ringrazio il *Village Garden*, perché correre dietro un pallone è ancora divertente.

Ringrazio le *Obese* perché un sorriso riescono sempre a tirarlo fuori. Anche più di uno.

Ringrazio *I Ragazzi* perché le serate/notti con voi sono state (e saranno) le migliori.

Infine ringrazio la mia famiglia, che mi ha permesso, non solo economicamente, di affrontare con serenità questo percorso lungo cinque anni.

Grazie di cuore a tutti.

Marco

Contents

Chapter 1

Introduction

In evolutionary biology the main purpose is to understand the history of life, studying the modification in the descendant, both in small-scale (changes in gene frequency in a population from one generation to the next) and in large-scale (the descendant of different species from a common ancestor over many generations). The central idea of biological evolution is that all life on Earth shares a common ancestor. Trying to come back to this primogenitor means to go through millennia, like using a time machine, in order to span the evolutionary process in a opposite direction.

The processes, which are involved in evolution and, nowadays, are known, are the transmission via sexual and asexual reproduction and a new method of sharing DNA: the so called **horizontal gene transfer (HGT)**. We define the horizontal gene transfer as new not in the sense that only from few time the organisms do it, but the scientists have dealt with this theory only since the 1950s. The horizontal gene transfer consists in the acquisition of genetic material from the environment, such as not only from the habitat where the organism lives, but also from the other organisms that share the same ecological niche. The protein-coding genes formed by HGT arise *de novo* in the whole DNA and this de novo gene birth and insertion in the DNA are still poorly understood.

With the growth of the volume of sequenced genomes and the bioinformatic methods to predict the protein synthesized by the genomes, the horizontal gene transfer's investigations are more precise. The majority of the system used to identify the presence of genes from other organisms are based on the analysis of the genome (or the protein synthesized) and the comparison with distant ones, in a phyletic way.

Whithin the biodiversity perspective, the ecological theories try to realize the within-tropic-level relations between the species in a community. A quantity that helps to understand this behaviour is the **relative species abundance (RSA)**. The relative species abundance in a community refers to how common or rare a species is relative to other species. Usually, the studies performed by a number of researchers have found that many species have a low number of individuals, while few species have many individuals.

Moreover there are a number of experimental observations and speculations on the

conceptualization of nucleic acids as a ecological system. Some genetic elements, as the transposons, are pieces of DNA capable to "parasitize" the nucleic acids, to jump from a position to another in the same molecule and to jump from one organism to another. Hence, according to the shape of experimental histograms and to the well known fact that nucleic acids are an ecosystem, we decide to describe the dynamics of protein domains by an ecological model. We apply a similar model, which Volkov et al. [52] used in describing the relative species abundance in a ecological community such as the coral reef.

The data are the counting of a protein domain in the sequenced bacterial genomes. We create a birth and death model with an immigration parameter, from which we obtain a binomial negative distribution that has been used to fit the experimental histograms. We link the value of the immigration parameter with the amount of horizontally transfer genes that a Family of bacteria has done: greater is this value, more horizontally transferred genes the Family has acquired.

Chapter 2

Biological background

In this chapter, we describe the principal biological notions, necessary in facing with the type of data as the ours. First we record the definition of protein domain and transmission of genetic material. Then we centre our discussion on the horizontal gene transfer and his studies. Finally we illustrate, neglecting the mathematical formalism, an analysis on the domains' families with the aim to emphasize the presence of a migration parameter.

2.1 Protein domain

A protein domain is defined as a polypeptide chain or a part of a polypeptide chain that can fold independently into a stable tertiary structure [7]. Domains vary in length from between about 25 amino acids up to 500 amino acids in length.

2.1.1 Three main classes of domains

On the basis of simple considerations of connected motifs, Michael Levitt and Cyrus Chotia have classified domain structures into three many groups [11].

This method is based on the secondary structures of the polypeptide chain, observed in the domain (Figure 2.2). The protein secondary structure is the general three-dimensional form of local segments of proteins (Figure 2.1). Secondary structure can be formally defined by the pattern of hydrogen bonds of the protein (such as α helices and β sheets) that are observed in an atomic-resolution structure.

α domains

The core is built up exclusively from α helices.

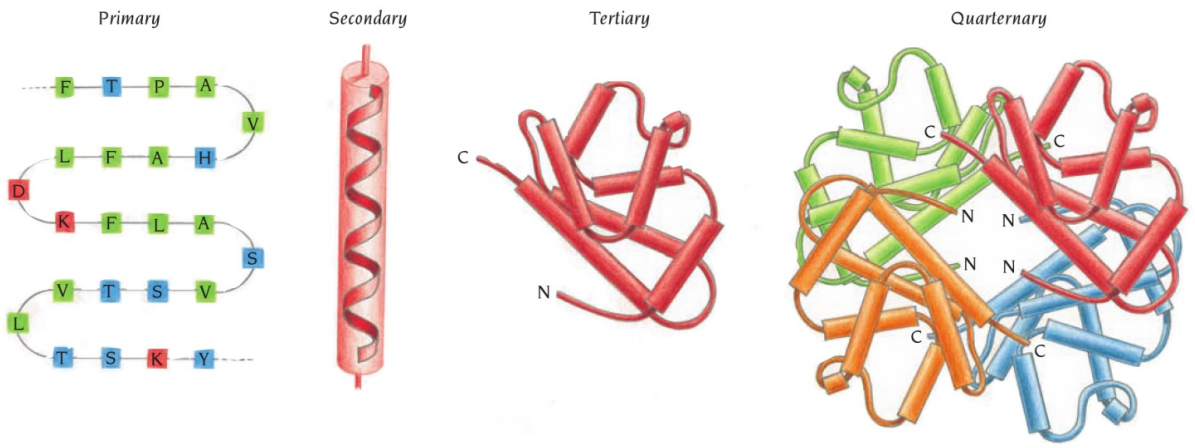

Figure 2.1: The primary structure of a protein is defined as the amino acido sequence of the polypeptide chain. Different regions of the sequence form local regular secondary structures. The tertiary structure is formed by packing such structural element. The final protein may contain several polypeptide chains arranged in a quaternary structure. [7]

β domains

The core comprises antiparallel β sheets and are usually two sheets packed against each other.

α/β domains

The structure is made from combinations of β-α-β motifs that form a predominantly parallel β sheet sorrounded by α helices.

2.1.2 Domains units of function

Domains are the fundamental units of tertiary structure of the protein and they are also units of function. There are many known examples where several biological functions that are carried out by separate polypeptide chains in one species are performed by domains of a single protein in another species.

For example, synthesis of fatty acids requires catalysis of seven different chemical reactions. In plant chloroplasts these reactions are catalyzed by seven different proteins, whereas in mammals they are performed by one polypeptide chain arranged in seven domains with short linker regions between the domains [47].

Moreover, domain recurrences among 3D structures consistently reveal that protein structure is more conserved than sequence [42]. There are many examples of domains adopting highly similar 3D structures despite no apparent similarity in sequence. For

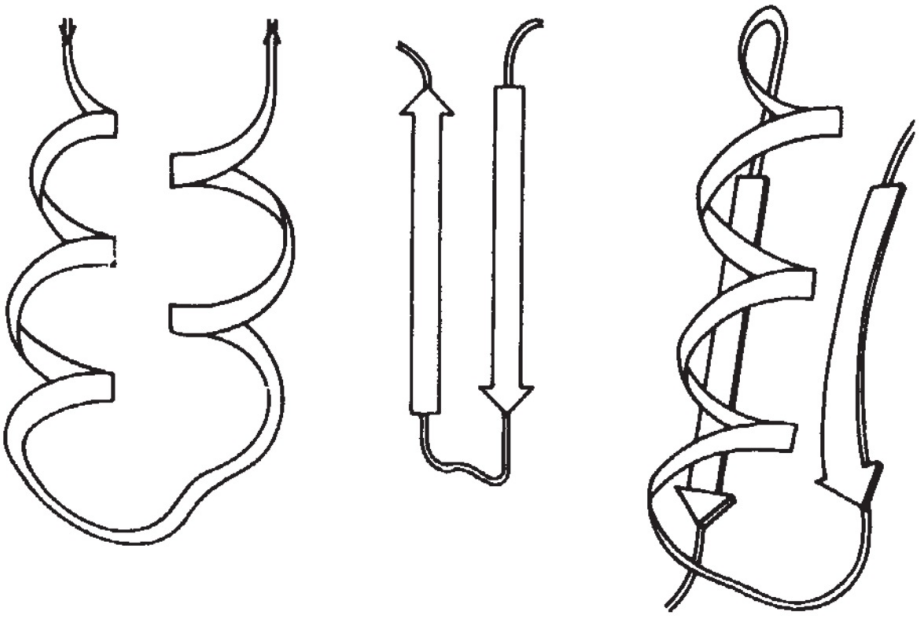

Figure 2.2: The three commonly occurring folding units: $\alpha\alpha$, $\beta\beta$ and $\beta\alpha\beta$. [10]

many of these examples, proteins have diverged beyond the limits of sequence similarity detection methods but have nevertheless retained a common structure and similar function. For example, adenylate cyclase and DNA polymerase contain a similar domain that was recognized by 3D structure comparison (Figure 2.3)[2].

2.2 Transmission of DNA

An organism can acquire new genetic material in two distinct and different way: vertical gene transfer and horizontal gene transfer. The fundamental difference between this processes is where the filament of DNA come from. In the next section is described quickly the vertical gene transfer, however the horizontal gene transfer occupies several pages.

2.2.1 Vertical gene transfer

In the traditional vertical transfer, the transmission of genes happens from parental generation to offspring via reproduction (sexual or asexual). In bacteria, the phenomena of vertical gene transfer is associated with the asexual reproduction, called binary fission (Figure 2.4).

While the chromosome is being duplicated, each copy starts to move to either poles of the cell. At the same time the parental cell elongates and grows. When duplication of

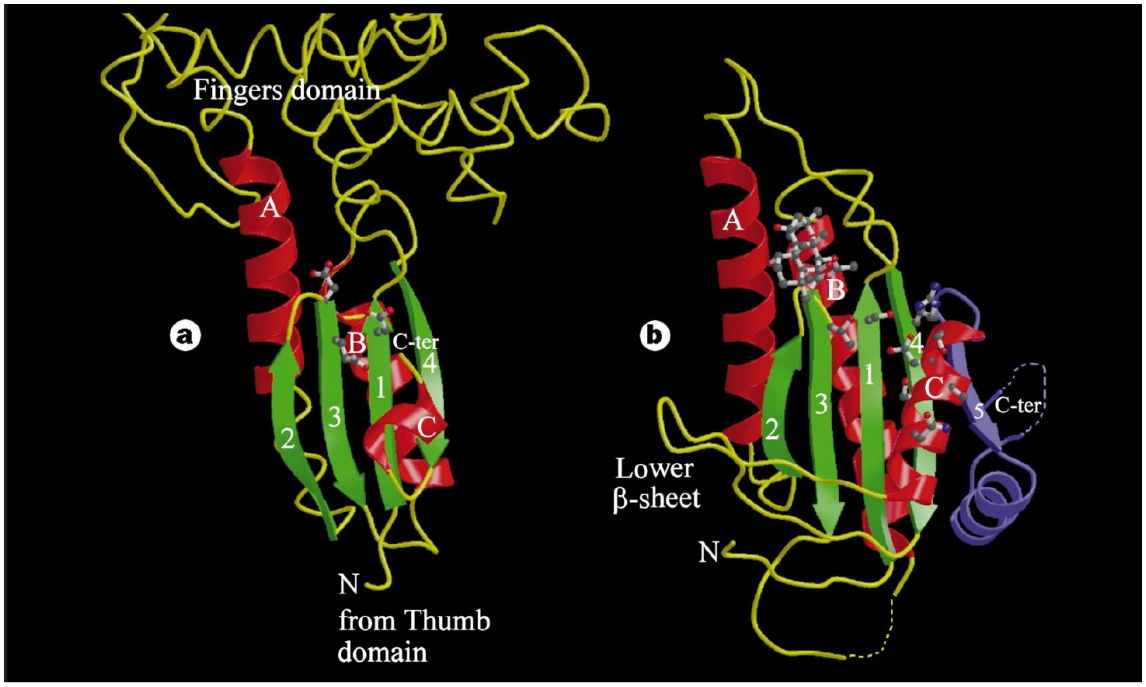

Figure 2.3: Chain of **a**, the palm domain of Thermus aquaticus polymerase and **b**, a monomer of adenylyl cyclase catalytic core. Equivalent helices and strands, shown as red coiled ribbons green arrows, respectively, occur in the same order in both structures. The similarity has important implications for the function and evolution of eukaryotic adenylyl cyclases and related proteins. [2]

the chromosome has been completed and the sizes of the bacterium are approximately doubled, the cell membrane folds inward, dividing in two identical cell daughter.

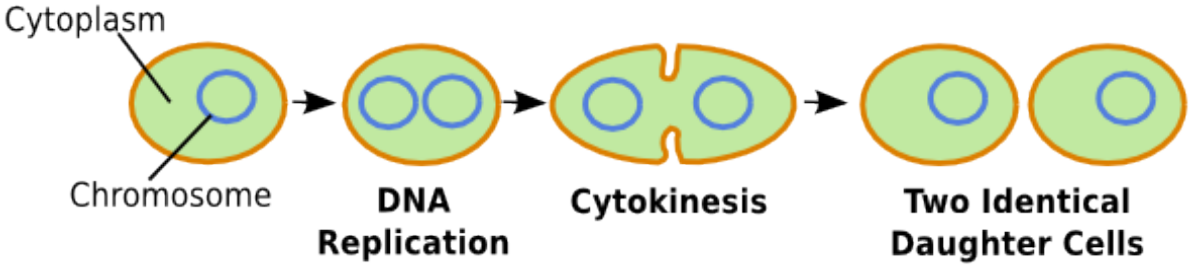

Figure 2.4: Scheme of the binary fission

2.3 Horizontal Gene Transfer

The Horizontal Gene Transfer (HGT), or lateral gene transfer, is an established way of evolution. Horizontal gene transfer is common among bacteria, even among very distantly related organisms.

The significance of horizontal gene transfer for bacterial evolution was not recognized until the 1950s, when multidrug resistance patterns emerged on a worldwide scale [16]. This process is thought to be a significant cause of increased drug resistance when one bacterial cell acquires resistance, and the resistance genes are transferred to other species. The facility with which certain bacteria developed resistance to the same spectrum of antibiotics indicated that these traits were being transferred among taxa, rather than being generated de novo by each lineage [41].

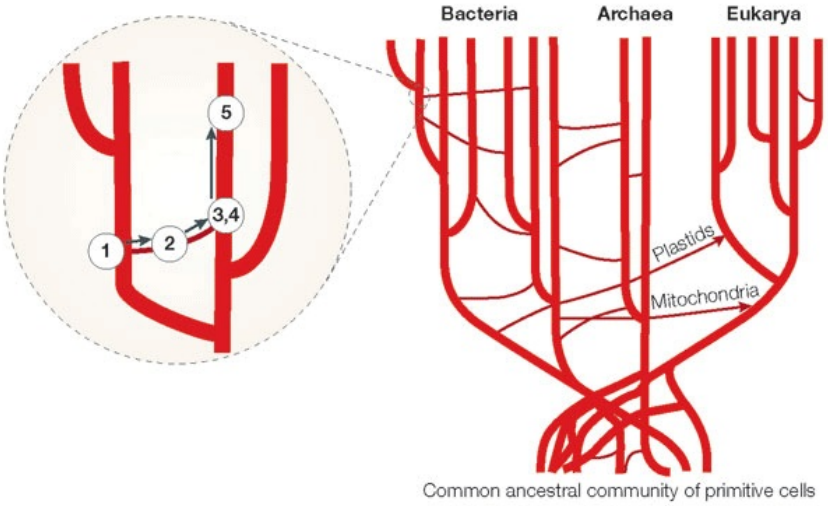

Figure 2.5: Scheme of horizontal gene transfer between distant branches taken from [46].

2.3.1 Mechanisms of transfer of DNA in HGT

In contrast to the evolution of new traits through the modification of existing sequences, the origin of new abilities through HGT has three requirements. First, there needs to be a means for the donor DNA to be delivered into the recipient cell. Second, the acquired sequences must be incorporated into the recipient's genome (or become associated with an autonomous replicating element). And third, the incorporated genes must be expressed in a manner that benefits the recipient microorganism. Observing the previous two conditions, the transfer of mobile genetic elements can occur by 3 methods: transduction, conjugation, and transformation (Figure 2.6).

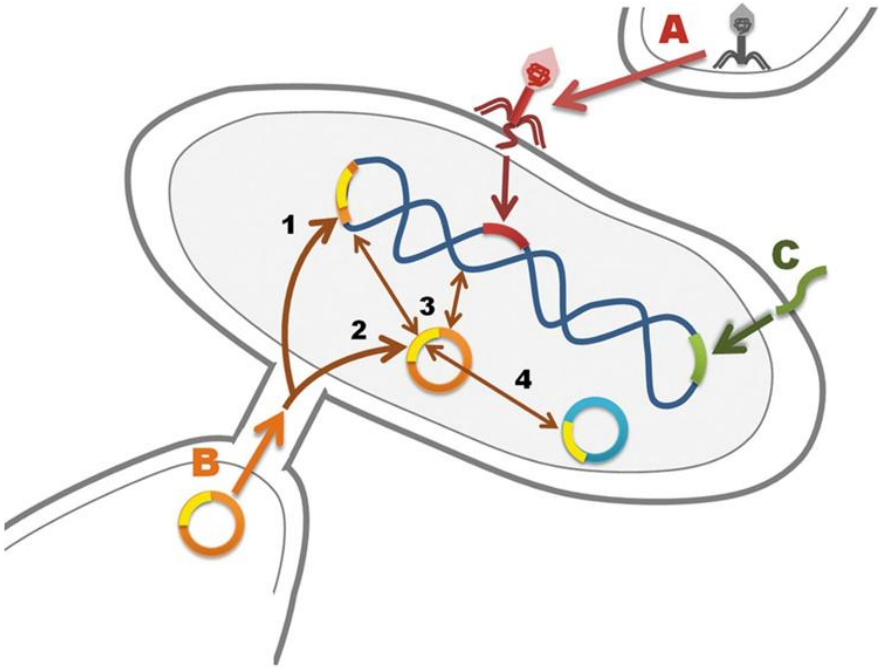

Figure 2.6: Mechanisms of DNA transfer between and within bacteria. **A**) Transduction: injection of DNA into a bacterium by a phage. **B**) Conjugation: plasmid in a donor bacterium is transferred through a pilus into a recipient bacterium; plasmid may integrate into the chromosome (1) or remain in the cytoplasm (2); plasmid may be transferred between cytoplasmic and chromosomal locations (3); plasmid may exchange insertion sequences or transposons with other plasmids (4) or the chromosome. **C**) Transformation: uptake of naked DNA from the environment.[24]

Transduction

In the process of transduction, bacterial genes are incorporated by bacteriophage particles and transferred to another bacterium. Transduction may be either *generalized*, whereby any bacterial gene may be transferred, or *specialized*, where only the DNA adjacent to the phage attachment site is transferred [17]. Bacteriophages have a restricted host spectrum, sometimes being limited to a single bacterial species. It depends upon microorganisms' receptors recognized by the bacteriophage. Furthermore, bacteria may mutate to become incapable of phage adsorption. The amount of DNA that can be transferred in a single event is limited by the size of the phage capsid, but can range upwards of 100 kilobases (kb) [41]. Transduction does not require donor and recipient cells to be present at the same place, or even the same time. On the other hand, phage-encoded proteins not only mediate the delivery of double-stranded DNA into the recipient cytoplasm, but can also promote the integration of DNA into the chromosome and protect the transferred sequences from degradation by host restriction endonucleases.

Conjugation

Conjugation involves physical contact between donor and recipient cells and can mediate the transfer of genetic material between domains. Typically, DNA is transferred from a donor to a recipient strain by either a self-transmissible or mobilizable plasmid. Conjugation can also mediate the transfer of chromosomal sequences by plasmids that integrate into the chromosome, and by conjugative transposons, which encode proteins required for their excision from the donor, formation of a conjugative bridge and transposition into the recipient strain.

Transformation

Transformation involves the uptake of naked DNA from the environment by bacteria that are in a state of natural competency, a physiologic state in which the bacteria are able to take up DNA and become transformed. Transformation has the potential to transmit DNA between very distantly related organisms. Single-stranded DNA is passed through the cell wall and cell membrane into the host through complex energy-requiring processes and enters the bacterial chromosome, mainly by homologous recombination [31] but also by the transient expression of nonhomologous recombination mechanisms encoded by the invading foreign DNA [18]. Some bacterial species or strains within a bacterial species are more prone than others to be naturally competent for foreign DNA uptake [34].

2.3.2 Criteria for detecting horizontally transferred genes

All criteria for identifying probable horizontal gene transfer, or more precisely acquisition of foreign genes by a particular genome, inevitably rely on some unusual feature(s) of subsets of genes that distinguishes them from the bulk of genes in the genome. Therefore all indications for horizontal transfer necessarily remain probabilistic, and the point of using different criteria is maximizing the likelihood of these events being identified correctly.

Unexpected ranking of sequence similarity among homologs

The suspicion of horizontal gene transfer usually emerges when a gene sequence from a particular organism shows the strongest similarity to a homolog from a distant taxon. The size of this fraction depends, evidently, on the genome and also on the cutoff (usually expressed in terms of alignment score or expect value) used to define “more similar” [32]. Generally the evidence from sequence comparisons should be considered preliminary. To make the case for horizontal transfer convincing, phylogenetic analysis is required.

Unexpected phylogenetic tree topology

Analysis of phylogenetic tree topologies is traditionally the principal means to decipher evolutionary scenarios, including horizontal transfer events [49]. It is unfortunate, however, that phylogenetic analysis does not offer such clear-cut solutions in all suspected cases of horizontal gene transfer. It is common knowledge that phylogenetic methods are prone to a variety of artifacts, perhaps the most notorious being long-branch attraction [36]. This phenomenon is particularly relevant for the analysis of probable horizontal gene transfer because these events may be accompanied by accelerated evolution, hence long branches in phylogenetic trees. Tree topology is a good indicator of the probable course of evolution only in cases when the critical nodes are strongly supported statistically, by bootstrap analysis or other methods [20]. On a more practical note, phylogenetic analysis is time and labor consuming, critically depends on correct sequence alignments, and is hard to automate without compromising the quality [32].

Unusual phyletic patterns

With many complete genome sequences available, new and relatively simple, but potentially powerful, approaches to evolutionary analysis become feasible. With the systematic delineation of families of orthologs (direct evolutionary counterparts related by vertical descent), the notion of a phyletic (phylogenetic) pattern has been introduced [50]. In the most straighforward formulation, a phyletic pattern is simply the pattern of species present or missing in the given cluster of orthologs. Certain types of phyletic patterns, however, appear to signal horizontal transfer in a more specific fashion.

Conservation of gene order between distant taxa

The evolution of bacterial and archaeal genomes involved extensive gene shuffling, and there is little conservation of gene order between distantly related genomes [28]. It has been determined that the presence of three or more genes in the same order in distant genomes is extremely unlikely unless these genes form an operon [54]. Therefore, when a (predicted) operon is present in only a few distantly related genomes, horizontal gene transfer seems to be the most likely scenario. If such cases can be confirmed by phylogenetic tree analysis for multiple genes comprising the operon, they figure among the strongest indications of horizontal transfer.

Anomalous nucleotide composition

Anomalous nucleotide composition is widely used but is applicable only to recent horizontal transfers. This approach is based on the "genome hypothesis," according to which codon usage is distinct signature of each genome [23]. Thus, genes whose nucleotide or

codon composition are significantly different from the mean for a given genome are considered as probable horizontal acquisitions although the likely source of these alien genes generally cannot be identified [37]. Many of the horizontally transferred genes revealed by these criteria are prophages, transposons, and other genetic elements for which such evolutionary mobility is not unexpected.

2.3.3 HGT in evolution

The fact that genes can move between distant branches of the tree of life (Figure 2.7) even at low probabilities raises challenges to scientists trying to reconstruct evolution by studying genes and gene sequences in different organisms, although it is hard to unequivocally determine which organism is the donor and which one is the receiver [32]. The logic used to formulate such hypotheses is based primarily on the "out of Africa" principle [8], which assumes that if HGT is occured, the taxon with the most diverse representation of the given family is the most likely source.

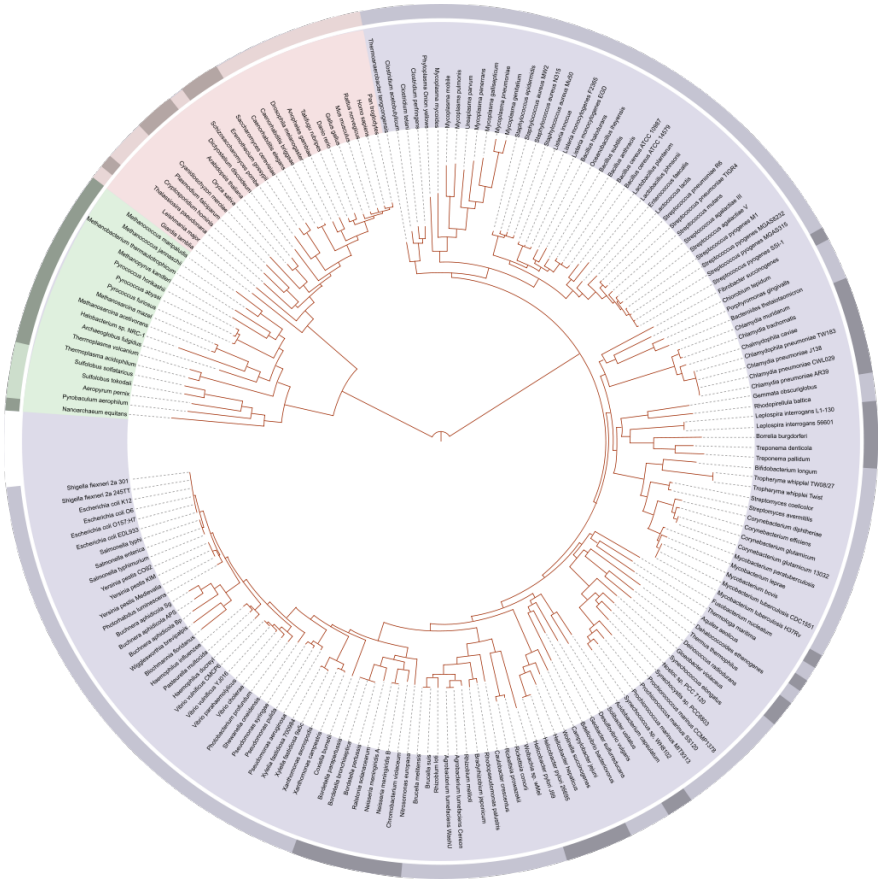

Figure 2.7: Modern representation of the Tree of Life, created from ITOL [33]

2.3.4 HGT in homo sapiens

Recent estimates suggesting that on average 81% of prokaryotic genes have been involved in HGT at some point [14]. However, relatively few cases have been documented in multicellular organisms [19], associated with few genes, remaining unclear the extent of horizontal gene transfer. Crisp et al. [13] carry out a detailed examination of HGT in 26 animal species, including the human genome. Now, we describe their method and results for the homo sapiens' case.

Their labour starts with the download of the trascriptome of each species. Every trascriptome is analysed with *blastx*, which is a type of searching of the bioinformatics tool BLAST (Basic Local Alignment Search Tool). Blastx identifies potential protein products encoded by a nucleotide query. This research is done against two databases: Metazoan (excluding phylum under analysis) and NonMetazoan. For every sequence of trascriptome is attached a bitscore for each top hit and is calculated the HGT index, h, as the difference between the bitscores of the best non-metazoan and the best metazoan matches. The HGT score h gives a relative quantitative measure of how well a given gene aligns to non-metazoan versus metazoan sequences, with positive numbers indicating a better alignment to non-metazoan sequences [6].

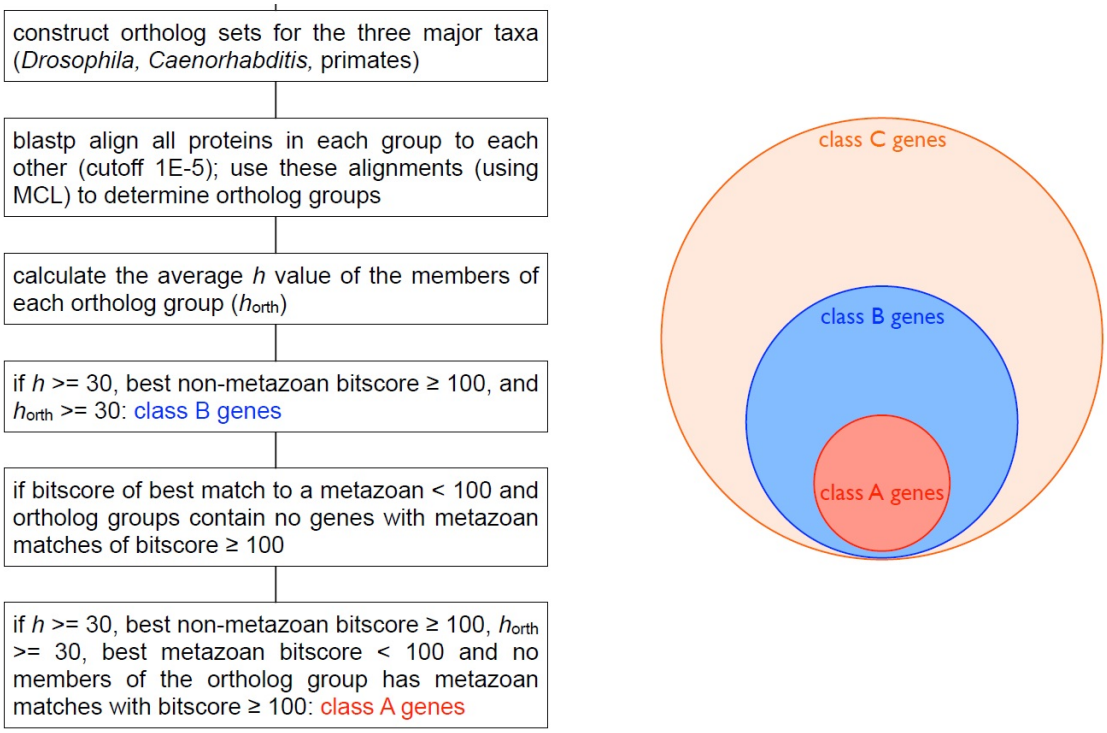

Figure 2.8: Steps followed for labeling genes. The image is taken from the supplementary material of [13].

The sequences are classified in two main group: class C, related with a base level of HGT, if $h \geq 30$ and the bestbest non-metazoan bitscore ≥ 100, and the native genes, for all the other genes. Afterwards, some genes, owned in class C, are classified in a subgroup, named class B, using taxon's information and building a new variable h_{orth}. Finally, it's applied a still more stringent filter to define class A foreign genes, a subset of class B, which have only very poor alignments to metazoan sequences. The more specific explanations of the method used for the creation of the classes are shown in the Figure 2.8.

Then it is performed phylogenetic analyses for all genes of each of the above classes and found that an average of 55% of all class C genes, 65% of all class B genes and 88% of all class A genes were phylogenetically validated as foreign. The first report of the human genome sequence highlighted 223 protein sequences that were proposed to originate from bacteria by HGT [12], but many were rejected as foreign [48]. Crisp et al. [13] identify up to 128 additional foreign genes in the human genome (128 class C, of which 93 are class B and 33 class A), giving a total of 145 class C genes, of which 110 are class B and 39 class A (Figure 2.9). In conclusion, HGT has contributed to the evolution of many, perhaps all, animals and that the process is ongoing in most lineages and the majority of these genes are concerned with metabolism, suggesting that HGT contributes to biochemical diversification during human evolution.

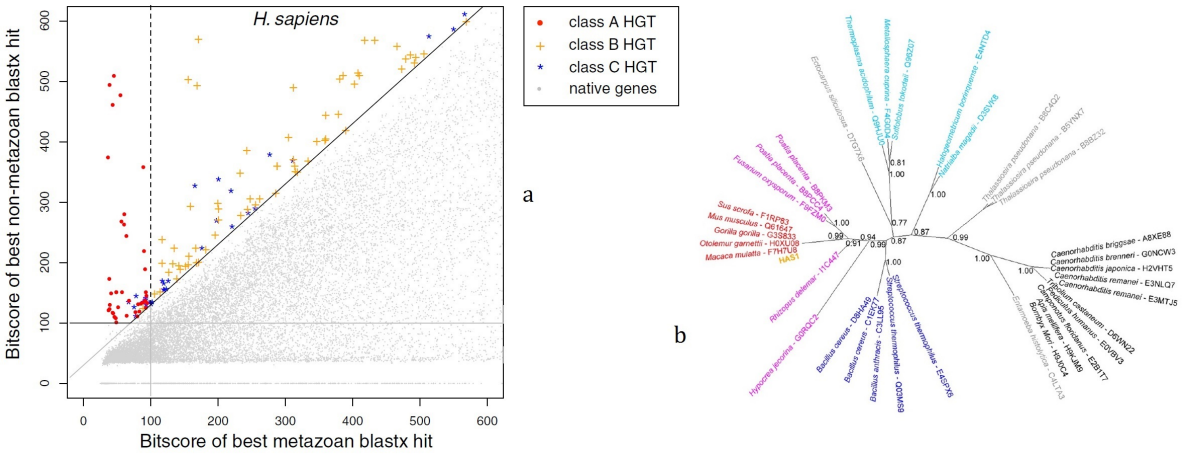

Figure 2.9: **a**) The panel shows the scores for all genes in H. sapiens, colour-coded according to their classification (class A: red, class B: orange, class C: blue, native genes: grey). **b**) Phylogenetic tree for the human gene HAS1 shows that this gene is found in a wide variety of species. For each branch the species name and UniProt accession is shown. The human gene under analysis is shown in orange, proteins from chordates are in red, other metazoa in black, fungi in pink, plants in green, protists in grey, archaea in light blue and bacteria in dark blue.

2.4 Birth-Death-Innovation Model (BDIM)

Karev et al. [30] present a birth and death model and apply it in the observed distributions of domain family size in diverse prokaryotic and eukaryotic genomes. A genome is treated as a *bag* of coding sequence for protein domains and each domain is considered to be a member of a family.

Three types of elementary evolutionary events are considered.

Domain birth which generates a new member within a family; the principal mechanism of birth is duplication with divergence but additional mechanisms may be considered, including acquisition of a family member from a different species via horizontal gene transfer.

Domain death which results from domain inactivation and/or deletion.

Domain innovation which generates a new family with one member; innovation may occur via horizontal gene transfer from another species, via domain evolution from a non-coding sequence or a sequence of a non-globular protein, or via major change of a domain from a pre-existing family after a duplication, which makes the relationship between the given domain and its family of origin undetectable.

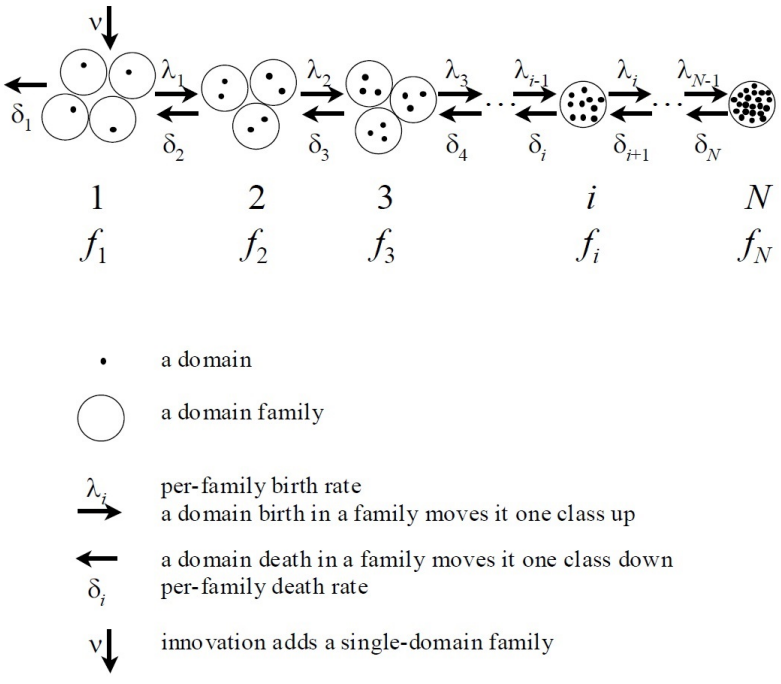

Figure 2.10: Domain dynamics and elementary evolutionary events under BDIM. [30]

The simplest model that resulted in a good fit to the observed domain family size distributions was the secondorder balanced linear BDIM, based on a model very close to the one which is presented in the section 3.5. In the Figure 2.11 are shown the fit of the protein domain family size distribution in the case of the human genome.

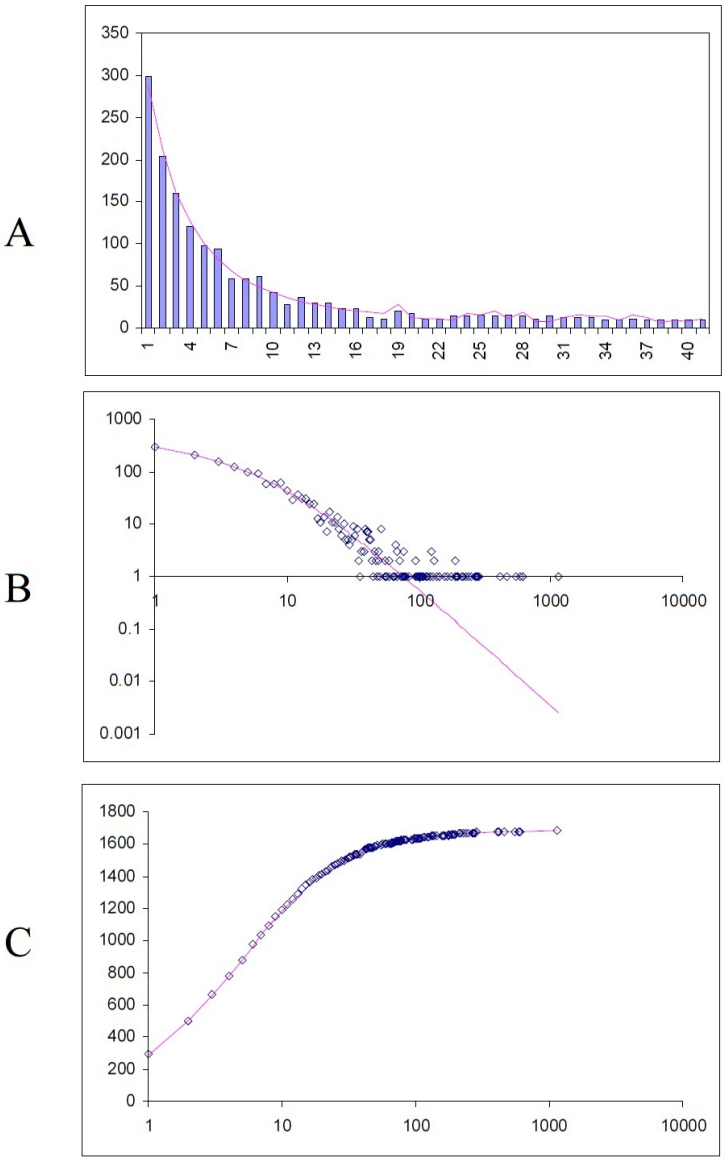

Figure 2.11: Fit of empirical domain family size distributions of the genome of Homo sapiens to the second-order balanced linear BDIM. **A**) Distribution of the size of domain families grouped into bins **B**) Domain family size distribution in double logarithmic coordinates. **C**) Cumulative distribution function of domain family size. [30]

Chapter 3

Neutral theory and stochastic model

In this chapter we are going to illustrate the main ecological theories, aiming to introduce the relative species abundance. Before describing the relative species abundance in inductive approaches, we focus on the difference between deterministic and stochastic model. Then we have a look at the birth and theory model, considering only models with continuous time although an analogous theory exists for stochastic processes with discrete time. Finally we propose the model used to fit our data, based on the theory of birth and death model and the ecological theory of neutrality.

3.1 Ecological theories

The main purpose of modern ecological theories is to describe and explain the within-trophic-level biodiversity [26]. In this contest, we refer the word biodiversity to the relative species abundance, which stands for the relation of common and rare species, and the species richness, which is the total number of species. The tropic level of a species is the position it occupies in the food chain, so on the organisms observed potentially or actually compete for similar resources. There are 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. For our purposes, 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.

In ecology, two main school of thought dominate the modern theory. As the controversy between determinism and stochasticity in modern physics, in ecology there are two conflicting world views on the nature of ecological communities: the niche and the dispersal perspectives.

3.1.1 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 [45]. In the ecological niche point of view, the species' interactions with their environment are defined by two elements [9]:

- the requirement for an organism of a given species to live in a given environment, the extent to which a limiting factor (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 (the availability of a resource or the density of a predator or parasite).

Several theories on species diversity have been developed around the notion of interspecific competition. All species are to some extent limited by their resources or natural enemies. In communities the species that is able to maintain a positive per capita growth rate at the lowest resource level or highest natural enemy pressure will drive all other competing species to extinction. This is the competitive exclusion principle [51].

Because of niche partitioning it is possible that the competing species coexist stably. This coexistence requires that any competing species, which are relatively rare, must have a higher growing potential than the other and more abundant competing species. 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 [45].

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 [26]. On the other hand, this theory is not able to predict a limit to diversity, and consequently neither to explain species relative abundance.

3.1.2 Neutral theory

The dispersal assembly perspective is opposed against the theory described before and it asserts that the communities are open, nonequilibrium assemblages of species largely thrown together by history, chance and random dispersal [26]. In this section, we focus on a class of dispersal conjecture, called neutral, in which all individuals of all species have equivalent per capita probabilities of giving birth, dying, migrating and speciating, such events occurring randomly for any given individual [51].

The neutral theory is defined at the individual level. All that is required is that all individuals of every species obey exactly the same rules of ecological engagement. The apparent stability of species diversity in the community can be attributed to a balance between speciation or immigration processes and the gradual loss of competing species diversity caused by demographic stochasticity (ecological drift) and competitive exclusion.

Neutral models are powerful because of their minimalist aspects, and because they predict a surprising number of complex patterns of competing species communities and metacommunities that might accurately describe species abundance and species relationships in the field [51].

3.2 Deterministic model vs stochastic model

A fundamental query in creating a new model is how much the randomness of the phenomena influences the physical magnitudes of the problem. To answer this question, there are two main class of type of model: deterministic and stochastic one.

In the deterministic models there is not randomness in evaluating the quantities both in the future both in the past. Usually, this models are ruled by differential equation and, setted once the initial conditions, the solution of the differential equation is unique and every magnitudes is evaluated.

In the stochastic models, contrariwise, the randomness plays a significant role in computing the output. Indeed, in this type of models only the probabilities density functions are known and, setting the initial conditions, the outputs rarely overlap in two different evaluation.

3.3 Relative Species Abundance

Relative species abundance (RSA) is a component of biodiversity and refers to how common or rare a species is relevant to other species in a defined location or community [26].

Observing the patterns of relative species abundance in diverse array of ecological communities (Figure 3.1), we can note all of them 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, nevertheless they have a curiously analogue fitness. 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.

In the following sections, we describe the two milestones of the deductive approach in the study of relative species abundance, which dominate in the earliest years. This methods start with fitting the observed distributions to statistical distributions and,

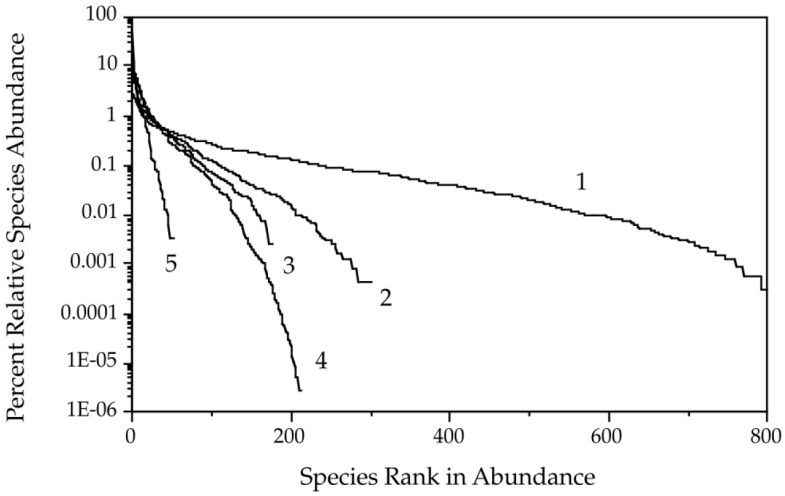

Figure 3.1: Patterns of relative species abundance in different ecological communities. Species in each community are ranked in percentage relative abundance from commonest (left) to rarest (right). (1) Tropical wet forest in Amazonia. (2) Tropical dry deciduous forest in Costa Rica. (3) Marine planktonic copepod community from the North Pacific gyre. (4) Terrestrial breeding birds of Britian. (5) Tropical bat community from Panama. [26]

afterwards, there are little attempts to explain theoretically what happens in the community, to reach the state.

3.3.1 Logseries distribution

Fisher analysed abundance data, from the labours of Corbet and Williams [22], about butterfly in Malaya and British moths. 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. The resulting compound distribution was negative binomial.

The problem that Fisher beat was the disability of counting the zero abundance class. The scientist trashed out truncating the negative binomial distribution and furthermore assumed that the total number of species in a community was infinite.

According to the logseries, as it is now generally called, he obtained a one parameter distribution and the number of species in a collection having n individuals will be given by

$$S_n = \alpha \frac{x^n}{n} \tag{3.1}$$

where x is a positive number lower than 1 and α is a measure of diversity.

Adding all terms, the total number of species, S, and the total number of individuals in the collection, N, are expected to be

$$S = \alpha \left[-\ln (1 - x) \right] \tag{3.2}$$

and

$$N = \alpha \frac{x}{1 - x} \, . \tag{3.3}$$

The parameter Fisher's α, is a widely used measure of species diversity because it is theoretically independent of sample size [22].

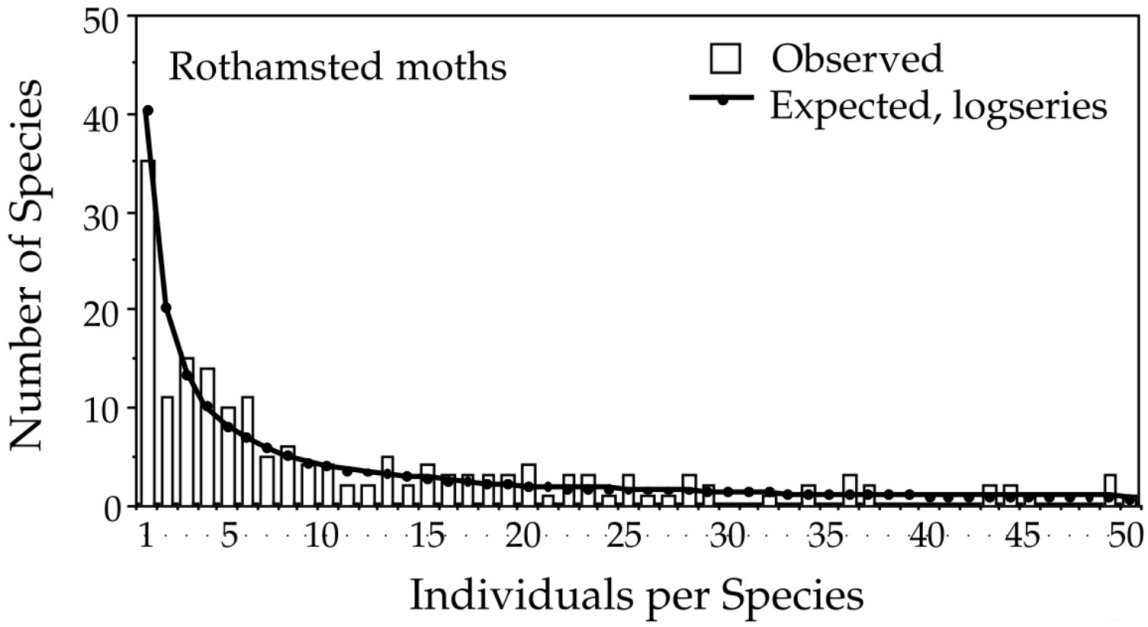

Figure 3.2: An example of the use of the logseries distribution to fit data on species abundance in collections of moths from [26].

3.3.2 Lognormal distribution

Preston criticized the logseries on the grounds that it was not a good fit to data that he had assembled, primarily on bird species abundances. He noted that the distributions had curves similar to bell-shaped ones, such that species having intermediate abundances were more frequent than very rare species. Preston observed that the distributions were lognormal and introduced a simple way to display this lognormal distribution of relative species abundance, based on the split of the category by the powers of 2. The process to create this histogram is described carefully in the section 3.3.3.

Otherwise the case of the logseries, in which the distributions are discrete, the lognormal distribution is continuous. 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 bin, then the so called Species Curve, in the Rth doubling abundance class, can be written as

$$S_R = S_0 e^{-(aR)^2} \tag{3.4}$$

where R is an integer and a is a constant that depends on the variance of the distribution, $a = 1/\sqrt{2\sigma}$. It is then possible to predict how many species are in the community by calculating the total area under the curve

$$N = S_0 \frac{\sqrt{\pi}}{a} \,. \tag{3.5}$$

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 [45].

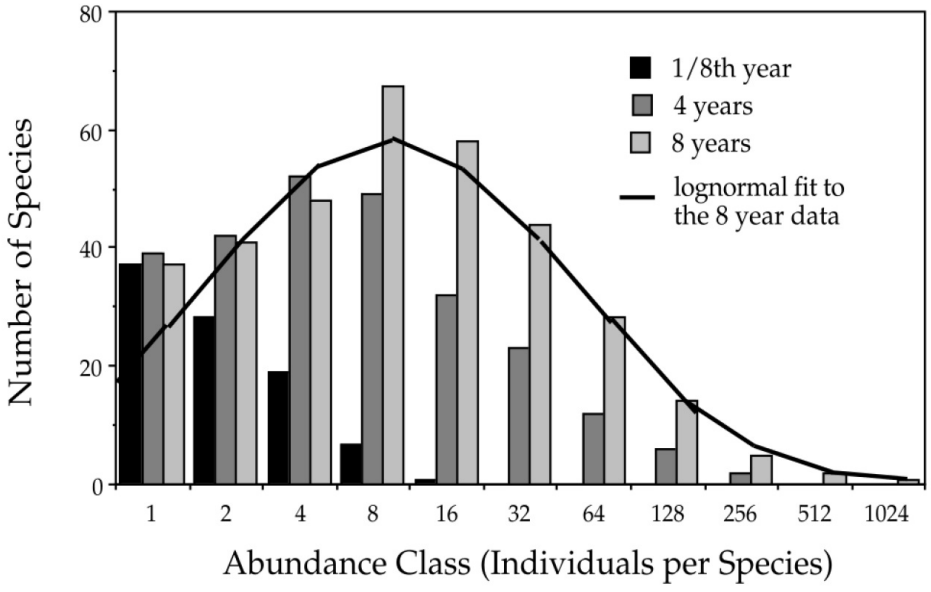

Figure 3.3: As the survey of moths at light traps at Rothamsted Field Station was extended over more years, the distribution of individuals per species became lognormal, as Preston predicted. [26]

3.3.3 Preston plot

Preston built doubling categories of abundance (1, 2, 4, 8, 16, etc.), and counted the species having abundances falling in each category (Figure 3.4). These groups are a sequence of octaves of frequency.

Species group	A	B	C	D	E	F	G	H	I	J	etc.
Approximate specimens observed of that species	1	2	4	8	16	32	64	128	256	512	etc.

Figure 3.4: Scheme proposed by Preston in his work [43].

An octave is simply an interval of two-to-one and if a species is represented by a boundary value, half species is credited in the category before and the second half in the next category.

In our work we regain the Preston's system, but with two alterations. First, we change the condition of a boundary species: in our labour we put it in the previous category. For example, the first category carries the singleton, the second contains the two occurrence species and so on. Second we fit our histograms with a binomial negative distribution, obtained from the model described in the section 3.5.

3.3.4 Dynamical model

We now present a simple unified theory for understanding these RSA patterns, developed by Azaele et al. [3] in the continuous form. 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.

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, one deterministic and the other stochastic.

The deterministic process has two contributions: an immigration rate b, which, for simplicity, is assumed to be equal for all species and independent of time and 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 accounted for by the deterministic part and is proportional to$\sqrt{x}$ from the central limit theorem.

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

$$\frac{dx}{dt} = a(x) + b(x)\xi(t)\,, \tag{3.6}$$

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

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}\left(\rho a(x)\right) + \frac{1}{2}\frac{\partial^2\left(b^2(x)\rho\right)}{\partial x^2}\,. \tag{3.7}$$

Thus the Langevin equation corresponding to this model is

$$\dot{x}(t) = b + \frac{x(t)}{\tau} + \sqrt{Dx(t)}\xi(t) \tag{3.8}$$

where $x > 0$ for any $t > 0$; b, τ and D are positive real constants, $\xi(t)$ is a Gaussian white noise with zero mean value and with time correlation $\langle \xi(t)\xi(t')\rangle = 2\delta(t-t')$.

The corresponding Fokker Planck equation for this process is

$$\dot{p} = \partial_x\left[\left(\frac{x}{\tau} - b\right)p\right] + D\partial_x^2\left(xp\right) \tag{3.9}$$

where $p = p(x,t)$ is the probability distribution function of finding x individuals at time t in the community. Accordingly, the fraction of species with a population between n and $n + \Delta n$ is $\int_n^{n+\Delta n} p(x,t)dx$. Setting $\dot{p} = 0$ we obtains the stationary solution of equation 3.9

$$p_0(x) = (D\tau)^{\frac{b}{D}}\,\Gamma^{-1}\left(\frac{b}{D}\right)x^{\frac{b}{D}-1}e^{\frac{x}{D\tau}}\,. \tag{3.10}$$

where $\Gamma(x)$ is the gamma function.

Furthermore $\dot{p} = 0$ in formula 3.9 implies that

$$\partial_x\left[\left(\frac{x}{\tau} - b\right)p\right] + D\partial_x^2\left(xp\right) = \partial_x\left\{\left[\left(\frac{x}{\tau} - b\right)p\right] + D\partial_x\left(xp\right)\right\} = 0\,. \tag{3.11}$$

The obvious solution is that the parenthesis is constant and, setting this constant equal to 0, we obtain

$$D\partial_x\left(xp\right) = \left(b - \frac{x}{\tau}\right)p\,. \tag{3.12}$$

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

$$\partial_x\left(Dxp\right) = \frac{b - \frac{x}{\tau}}{Dx}Dxp\,, \tag{3.13}$$

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.

$$p_0(x) = \frac{1}{Dx}e^{\int \frac{b-x/\tau}{Dx}dx} = \frac{1}{Dx}e^{\frac{b}{D}\ln x - \frac{x}{D\tau}} = \frac{1}{D}x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}} \tag{3.14}$$

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

$$\frac{1}{D}\int_0^\infty x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}}dx = 1\,, \tag{3.15}$$

that we can rewrite as

$$\frac{1}{D}\left(D\tau\right)^{\frac{b}{D}-1}\int_0^\infty \frac{x^{\frac{b}{D}-1}}{\left(D\tau\right)^{\frac{b}{D}-1}}e^{-\frac{x}{D\tau}}dx = 1 \tag{3.16}$$

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

$$\frac{1}{D}\left(D\tau\right)^{\frac{b}{D}-1}\int_0^\infty t^{\frac{b}{D}-1}e^{-t}D\tau dx = 1\,. \tag{3.17}$$

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

$$\frac{1}{D}\left(D\tau\right)^{\frac{b}{D}}\Gamma(b/D) = 1 \qquad \Rightarrow \qquad \frac{1}{D} = \frac{\left(D\tau\right)^{-\frac{b}{D}}}{\Gamma(b/D)}\,, \tag{3.18}$$

and the stationary solution (equation 3.14) becomes

$$p_0(x) = P_{RSA} = \frac{\left(D\tau\right)^{-\frac{b}{D}}}{\Gamma(b/D)}x^{\frac{b}{D}-1}e^{-\frac{x}{D\tau}} \tag{3.19}$$

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.

This three parameters have an important ecological sense: τ is the characteristic timescale associated with 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 takes into account density dependence effects arising from immigration and/or speciation; and D accounts for demographic stochasticity.

3.4 Theory of birth and death processes in biology

This theory was developed in the beginning of the twentieth century as a result of attempts to model growth of a population, taking into account stochastic demographic factors. Importantly the simplest process provides a natural and useful theoretical framework for several areas of modern biology, such as estimation of the age of alleles, reconstruction of phylogenies and modeling various aspects of genome evolution.

A birth-and-death process is a stochastic process in which jumps from a particular state (number of individuals, cells, lineages, etc.) are only allowed to neighboring states:

$$0 \underset{r_1}{\overset{g_0}{\rightleftharpoons}} 1 \underset{r_2}{\overset{g_1}{\rightleftharpoons}} \cdots \underset{r_{n-1}}{\overset{g_{n-2}}{\rightleftharpoons}} n-1 \underset{r_n}{\overset{g_{n-1}}{\rightleftharpoons}} n \underset{r_{n+1}}{\overset{g_n}{\rightleftharpoons}} n+1 \underset{r_{n+2}}{\overset{g_{n+1}}{\rightleftharpoons}} \cdots \underset{r_N}{\overset{g_{N-1}}{\rightleftharpoons}} N \underset{r_{N+1}}{\overset{g_N}{\rightleftharpoons}} \cdots$$

This property considerably simplifies the mathematical analysis, but the process remains applicable to numerous real-world systems. etc. The results obtained with birth-and-death models can be compared with empirical data allowing one to either reject some of the initial assumptions, or accept the model as a useful tool for analysis and prediction of properties of the real system. In biology, the stochastic models are more realistic than the deterministic ones because counts of individuals are discrete by definition.

The general study of temporally continuous, stochastic models of population growth apparently started with the work of Feller [21]. The cardinal assumption was that the growth of a population can be represented by a Markov process. The state of the population at time t can be described by the value of a random variable $X(t)$ with the property,

$$Pr\left\{X(t) = n | X(t_0) = m_0, X(\tau_1) = m_1, \ldots, X(\tau_k) = m_k\right\} = \\ = Pr\left\{X(t) = n | X(t_0) = m_0\right\} , \quad (3.20)$$

for all $\tau_i \leq t_0$ and whenever $t_o < t$.

If we interpret $X(t)$ as a population size, then a birth-and-death process is a Markov process $X(t), t \geq 0$ such that, in an interval $(t, t+\Delta t)$, each individual in the population has the probability $g_n \Delta t + o(t)$ of giving birth to a new individual (probability of transition from state n to state $n+1$) and the probability $r_n \Delta t + o(t)$ of dying (probability of transition from state n to state $n-1$). The parameters g_n and r_n are called the birth rate and death rate, respectively, where n is the population size. Functionally, the application of the theory of birth and death processes consists of two stages: first, rates g_n and r_n have to be specified, and second, the resulting process, which depends on the parameters of the biological system, has to be analyzed.

The state probabilities $p_n(t) = Pr\left\{X(t) = n\right\}$ of the process being in state n at time t satisfies the following system of differential equations, called Kolmogorov forward equations [4]:

$$\frac{\mathrm{d}P_{n,k}(t)}{\mathrm{d}t} = P_{n-1,k}(t) g_{n-1,k} + P_{n+1,k}(t) r_{n+1,k} - P_{n,k}(t) \left(g_{n,k} + r_{n,k}\right) , \quad (3.21)$$

setting $g_{-1,k} = r_{0,k} = 0$, so the state space consists of non-negative integers. Generally, there are two types of random processes: one in which there are no restrictions on the allowed set of states and the other in which there are restrictions in the sense that some states have special properties. There are two types of special state: absorbing state and reflecting state. The absorbing state appears when once the process reaches this state, it is trapped forever. On the other hand, once the process reaches a reflecting state, it must return to the previously occupied one.

3.4.1 Examples of processes

The simplest case when the solution of equation 3.21 is straightforward is a pure birth process or the Poisson process. In this case, the rates are set $g_n = \lambda$ and $r_n = 0$, and the solution of 3.21 subject to the initial condition $p_0(0) = 1$ is the Poisson distribution

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \,. \tag{3.22}$$

It is well known that the distribution of the time intervals between any two successive jumps in any Markov process with continuous time and discrete space of states is exponential [1].

Another example, for a simple birth process with the initial condition $p_m(0) = 1$, it can be shown that the state probabilities are

$$p_n(t) = \binom{n-1}{m-1} e^{-\lambda m t} \left(1 - e^{-\lambda t}\right)^{n-m}, \quad n \geq m \,. \tag{3.23}$$

This stochastic process was first studied by Yule [55]. The state of the process was thought of as a species within a genus, and the creation of a new species by mutation was conceived as being a random event with the probability proportional to the number of species [40]. Yule used this process to explain the observed power law distribution of genera of plants having n species.

For a simple birth-and-death process, putting $g_n = \lambda$, $r_n = \mu$ and $p_1(0) = 1$, the solution of equation 3.21 is

$$p_0(t) = P_0 = \frac{\mu \left(e^{(\lambda-\mu)t} - 1\right)}{\lambda e^{(\lambda-\mu)t} - \mu} \tag{3.24}$$

$$p_n(t) = (1 - P_0) \left(\frac{1 - \lambda P_0}{\mu}\right) \left(\frac{\lambda P_0}{\mu}\right)^{n-1}, \quad n \geq 1 \,. \tag{3.25}$$

For other possible initial conditions, the solution of 3.21 is more complicated but still can be obtained.

3.4.2 Moran model

Let us consider a population of haploid individuals of a fixed constant size N. Let us further consider one locus that can have two types of alleles, A and a. Because of random drift (stochastic nature of births and deaths), one of alleles eventually goes extinct; if there are no mutations, the population becomes homogeneous and the interesting quantity is the speed with which it approaches homozygosity. Another situation is appearance of a unique mutant (e.g. A); in this case, the probability of fixation of A and the mean time of fixation are of particular importance. If we assume that mutations can go in

either direction, then the population will be heterozygous forever, and the variable of interest is the stationary distribution.

Two such models have been the basis of most of the work in population genetics: the Wright–Fisher model [22] and the Moran model [35]. The Wright–Fisher model describes populations with discrete, seasonal reproduction and non-overlapping generations, whereas the Moran model is most applicable to populations with continuous reproduction. The Moran model is important for two reasons: first, in contrast to the Wright–Fisher model, it applies to organisms with overlapping generations. Second, many results that can be obtained only approximately under the Wright–Fisher model can be derived exactly using the Moran model [40].

In order to analyze the model, we need to define the birth and death rates. The biological system can be described using several parameters: the population size, the current number of individuals that carry allele A, the selective advantage of A over a (or vice versa) and the mutation rates from A to a and from a to A.

Let there be n copies of allele A and $N-n$ copies of allele a. It is assumed that individuals carrying A have the selection coefficient s. The transition rates for the Moran model can be written as

$$g_n = (1+s)\,\frac{N-n}{N} p_n\,, \qquad r_n = \frac{n}{N}\,(1-p_n)\;, \tag{3.26}$$

where p_n is the probability that the choice results in an A if there are n copies of this allele in the population. Assuming that there are no mutations, $p_n = n/N$. If we assume that the mutation rate from A to a is v, and from a to A is u, then

$$p_n = \frac{n}{N}\,(1-v) + \frac{N-n}{N} u\,. \tag{3.27}$$

If $v \neq 0$ and $u \neq 0$, we deal with a birth and death process with reflecting boundaries ($g_0 \neq 0, r_N \neq 0$). The presence of reflecting boundaries means that there exists a stationary distribution p^* which is easy to calculate numerically noting that, at equilibrium, $p_n^* r_n = p_{n-1}^* g_{n-1}$ must be satisfied. When N is large, a good approximation for the stationary distribution can be found.

If we assume that there are no mutations, then ($g_0 = r_N = 0$), and we have a birth and death process with absorbing boundaries. One of the main questions in this case is the probability that a new mutant penetrates the entire population, called probability of fixation. In mathematical terms, this can be expressed as the probability of reaching the absorbing state N before reaching the absorbing state 0. The probability that the system ends up in the state N (the probability of fixation) if initially there is only one A is

$$P_{fix} = \frac{1}{1 + \sum_{i=1}^{N-1} \prod_{n=1}^{i} \frac{r_n}{g_n}}\,. \tag{3.28}$$

Inserting the formula 3.26 with $v = u = 0$ is readily evaluated to

$$P_{fix} = \frac{1-(1+s)^{-1}}{1-(1+s)^{-N}} \approx \frac{1-e^{-s}}{1-e^{-sN}}\,. \tag{3.29}$$

If $s = 0$, one can find that $P_{fix} = 1/N$, the probability of fixation of a neutral mutant.

3.4.3 Logistic growth

This model accounts for the density dependence in the growth of a single population. It is based on the hypothesis that the net birth rate per individual (i.e. the difference between the birth rate and the death rate) is a linearly decreasing function of the population size. This implies that the net population birth rate is a quadratic function of the population size. The model is closed in the sense that no immiagration or emigration is allowed. Mathematically, the deterministic logistic model leads to a nonlinear differential equation that can be solved explicitly

$$\frac{dN(t)}{dt} = mN(t)\left(1-\frac{N(t)}{K}\right) \qquad N(0) = N_0\,, \tag{3.30}$$

where $N(t)$ is the size of the population at moment t, $m > 0$ is the intrinsic growth rate, and $K > 0$ is the carrying capacity. All solutions of equation 3.30 monotonically lead to the asymptotically stable equilibrium $N^* = N(\infty) = K$.

The logistic stochastic process is important for several reasons. It is well appreciated that the genetic makeup of a population strongly depends on the population structure while most of the population evolution models assume a constant population size. Density-dependent effects influence the size of the population, preventing indefinite growth, and the logistic model is the simplest stochastic model with changing population size and density-dependent mechanisms that affect this size.

In the logistic model, the state zero is usually an absorbing state such that eventual absorption at the origin is certain, and all states except the origin are transient. The state is called transient if the process visits this state only finitely many times. The immediate two issues to address are the calculation of the mean time to extinction and the possible behavior of the system prior to extinction. However, the time to extinction may not have a known distribution, because of which characterizing the system by the mean time to extinction can be misleading. The behavior of the process prior to extinction can be productively explored within the framework of the so-called quasi-stationary distributions [15]. The quasi-stationary distribution cannot be found analytically but there are effective numerical methods for determining such distributions [38].

3.5 Our model

The model used is a generalized stochastic model that describes a birth and death process, with the addiction of a migration parameter. This model is very close to one used by Volkov et al. [52] for describing the relative species abundance in the description of the species-rich communities such as coral reefs.

We define $P_{n,k}(t)$ as the probability at the time t to find n appearances in the kth protein domain. Consequently the time-evolution of the probability is given by the master equation

$$\frac{\mathrm{d}P_{n,k}(t)}{\mathrm{d}t} = P_{n-1,k}(t)g_{n-1,k} + P_{n+1,k}(t)r_{n+1,k} - P_{n,k}(t)\left(g_{n,k} + r_{n,k}\right) , \tag{3.31}$$

where $g_{n,k}$ and $r_{n,k}$ correspond to, respectively, the probabilities of birth and death of the kth domain with n appearances, setting $g_{-1,k} = r_{0,k} = 0$. We obtain from the master equation the equilibrium solution

$$P_{n,k} = P_{0,k}\prod_{i=1}^{n}\frac{g_{i-1,k}}{r_{i,k}} . \tag{3.32}$$

We suppose a simple condition of the probability of birth and death of the kth domain:

$$g_{x,k} = B_k\left(S_k + x\right) \tag{3.33}$$

and

$$r_{x,k} = D_k x \tag{3.34}$$

where B_k and D_k represent the birth and death rates independent from density and S_k stands for the migration parameter.

Inserting the conditions 3.33 and 3.34 in the solution of the stationary state of the master equation (3.32), we obtain that the probability $P_{n,k}$ can be written

$$\begin{aligned} P_{n,k} = P_{0,k}\prod_{i=1}^{n}\frac{B_k\left(S_k + i - 1\right)}{D_k i} = P_{0,k}\prod_{i=1}^{n}\frac{B_k}{D_k i}\left(S_k + i - 1\right) = \\ = P_{0,k}\left(\frac{B_k}{D_k}\right)^n \cdot \frac{\left[\left(S_k\right)\left(S_k+1\right)\cdots\left(S_k+n\right)\right]}{n!} = \\ = P_{0,k}\left(\frac{B_k}{D_k}\right)^n \frac{\Gamma\left(n+S_k\right)}{n!\,\Gamma\left(S_k\right)} . \end{aligned} \tag{3.35}$$

Using the normalization condition, we compute

$$P_{0,k} = 1 + \sum_{i=1}^{\infty} P_{i,k} = \left(1 - \frac{B_k}{D_k}\right)^{-S_k} . \tag{3.36}$$

We concern that the probability $P_{n,k}$, putting in the normalization condition, follows a negative binomial distribution:

$$P_{n,k} = \left(\frac{B_k}{D_k}\right)^n \left(1 - \frac{B_k}{D_k}\right)^{S_k} \frac{\Gamma\left(n + S_k\right)}{n!\,\Gamma\left(S_k\right)}\,. \tag{3.37}$$

3.5.1 Distribution of protein domains

The number of domains which contains n appearances is given by

$$\varphi_n = \sum_{k=1}^{N_{Sp}} I_{n,k} \tag{3.38}$$

where N_{Sp} is the total number of the domains that may be in the genome and $I_{n,k}$ is a random value which takes value 1, with probability $P_{n,k}$, and 0, with probability $(1 - P_{n,k})$. So the average number of domains is given by

$$\langle\varphi_n\rangle = \sum_{k=1}^{N_{Sp}} P_{n,k} \tag{3.39}$$

Another useful parameter is the average number of observed domains in a genome:

$$N_{OBS} = \langle N_{Sp} - \varphi_0\rangle = N_{Sp} - \sum_{k=1}^{N_{Sp}} \left(1 - \frac{B_k}{D_k}\right)^{S_k}\,. \tag{3.40}$$

Using the hypothesis of neutral ecological equivalence $S_k = S$, $B_k = B$ and $D_k = D$ are the same for all domains. This assumption involves writing the formula 3.39, putting in the equation 3.37, as

$$\langle\varphi_n\rangle = N_{Sp} \cdot P_n = N_{Sp} \cdot \left(\frac{B}{D}\right)^n \left(1 - \frac{B}{D}\right)^{S} \frac{\Gamma\left(n + S\right)}{n!\,\Gamma(S)}\,. \tag{3.41}$$

In addiction we obtain the value of the total number of protein domains that may be present in the community inverting the equation 3.40:

$$N_{OBS} = N_{Sp} - N_{Sp} \cdot \left(1 - \frac{B}{D}\right)^{S} \quad \Rightarrow \quad N_{Sp} = \frac{N_{OBS}}{1 - \left(1 - \frac{B}{D}\right)^{S}}\,. \tag{3.42}$$

Putting the value of N_{Sp} in the formula 3.41, we obtain the average number of domains with n appearances, in function of the number of observed domains and the parameters

$\frac{B}{D}$ and S:

$$\langle \varphi_n \rangle = \frac{N_{OBS}}{1-\left(1-\frac{B}{D}\right)^S} \cdot \left(\frac{B}{D}\right)^n \left(1-\frac{B}{D}\right)^S \frac{\Gamma\left(n+S\right)}{n!\,\Gamma(S)} = \\ = \frac{N_{OBS}}{\left[\left(1-\frac{B}{D}\right)^{-S}-1\right]} \left(\frac{B}{D}\right)^n \frac{\Gamma\left(n+S\right)}{n!\,\Gamma(S)}. \qquad (3.43)$$

Chapter 4

Experimental method

4.1 Data set

Our data set comes from the Nederland and is formed by the number of appearances of a protein domain in the sequenced bacterial genomes . Our colleagues took the genomes from the *NCBI* database, then predict the proteins using *Prodigal* and, finally, the proteins are analyzed by *InterProScan* for predicting protein domains. In the Figure 4.1 it's how the data appear to us, visualizing as a matrix in an Excel's sheet, with nominal features end numeric features (the most are the appearances of a protein domain in a genome).

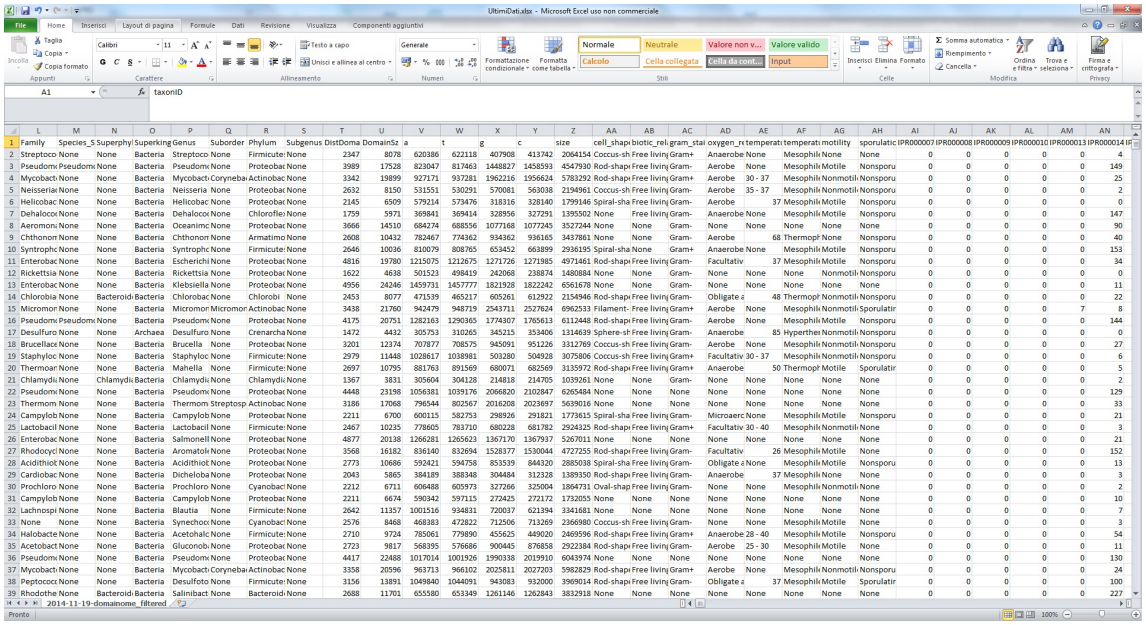

Figure 4.1: A little glimpse of the data set, using Microsoft Excel

4.1.1 Prodigal

Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) is a gene prediction algorithm. The algorithm is an open source program and, comparing to other gene-finding methods, has improved gene structure prediction, improved translation initiation site recognition and reduced false positives [27].

The advantages of this algorithm are the short time of evaluation (maximum thirty seconds per bacterial genome) and the absence of required training data. The algorithm's pseudocode is quoted in the Figure 4.2.

```
1.  Read in the sequence
2.  Locate all starts and stops in the genome
3.  Scan all open reading frames and record numbers of G's and C's in each codon position
4.  Build a frame bias model based on ORF length and G/C codon position within each ORF
5.  Record the highest scoring start nodes in each frame that overlap a stop codon by <= 60 bp
6.  Do the first pass dynamic programming, connecting nodes based on frame bias scores
7.  Create a hexamer background of all 6-mers in the entire sequence
8.  FOR each gene model in the dynamic programming output:
      1.  Gather all hexamer statistics
9.  Create log table of hexamer coding scores
10. FOR each gene model in the dynamic programming output:
      1.  Calculate a coding score based on hexamer statistics
      2.  Penalize the score if there is a higher scoring start upstream in the same ORF
      3.  IF the gene is very long but has a negative score, THEN give it a barely positive
          score
11. FOR 10 iterations
      1.  Build a ribosomal binding site and ATG/GTG/TTG background for all nodes
      2.  FOR each gene with a score of > 35.0:
            1.  Gather its Shine-Dalgarno RBS motif data and ATG/GTG/TTG data
      3.  Modify RBS and ATG/GTG/TTG weights by the observations
12. IF organism is not determined to use Shine-Dalgarno THEN run the non-SD finder
13. FOR each gene model:
      1.  Assign a final score of start score + coding score
      2.  Penalize the final score of genes < 250bp
14. Do the second pass dynamic programming, connecting nodes based on hexamer coding
15. FOR each gene model in the final dynamic programming:
      1.  Eliminate negative scoring models
      2.  Resolve very close start pairs (<= 15 bp from each other)
16. Print final output
```

Figure 4.2: Pseudocode description of the Prodigal algorithm [27]

4.1.2 InterProScan

InterProScan is a tool that scans the given protein sequences against the protein signatures of databases and it has a modular Java-based architecture (Figure 4.3).

Before InterProScan launches each of the protein sequence analysis applications, it takes advantage of pre-computed results whenever possible. It calculates a checksum for the query sequence and compares it with the checksums of the protein sequences that are present in a database called IPRMATCHES. When a match is found a IPR code is assigned to it which stands for a specific biological context. If the checksum calculated

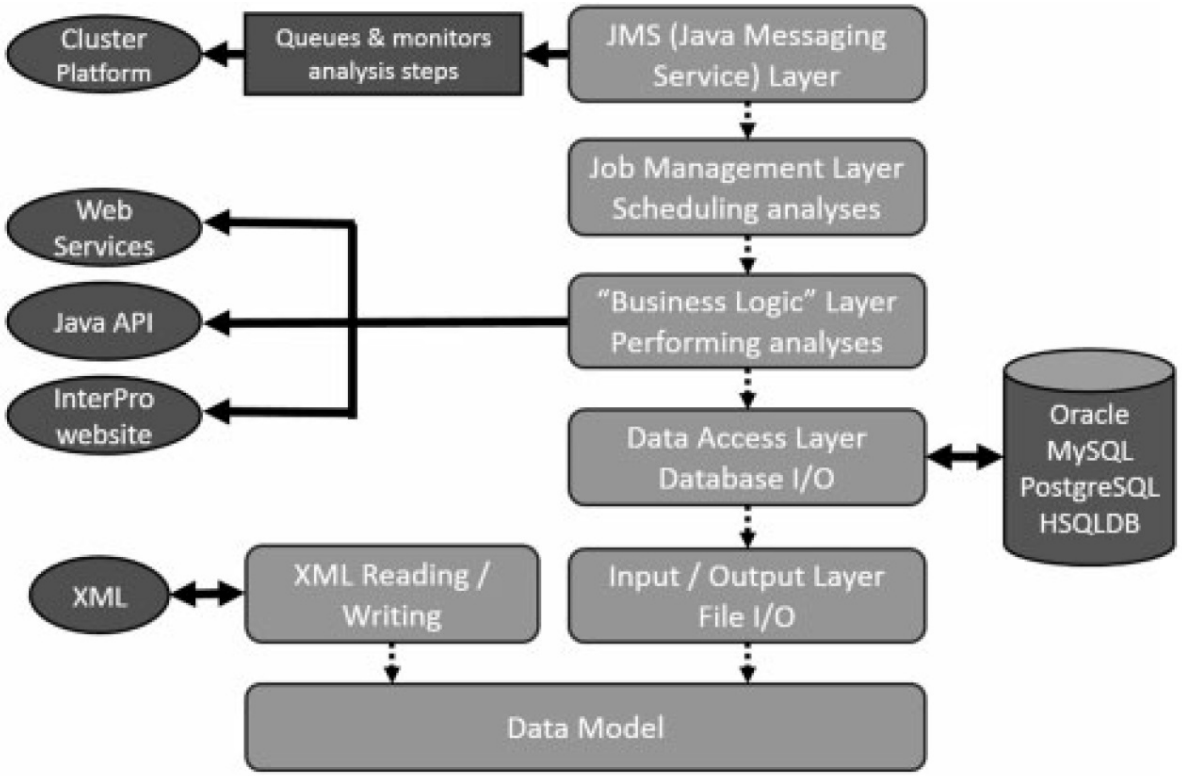

Figure 4.3: Overall system architecture of InterProScan [29]

for the query sequence does not match any checksums found in the IPRMATCHES database, the protein sequence analysis applications are launched in parallel [44].

For our goal, the algorithm used mostly Hidden-Markov-Model [5]. A Hidden-Markov-Model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. The outputs are strongly dependent on the unobserved states. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and, obviously, bioinformatics.

4.2 Fit

For our purpose, we need to create the preston plot of the protein domains of both every organisms and Family class. Then the histograms obtained are fitted and the parameters' values are estimated, using the numerical computing environment MATLAB. Finally, we analyse the results and test them creating a null model from our data.

4.2.1 Creation of Preston Plot

We use only the organisms which belong in a Family that has more than three genomes sequenced. The intention of this approach is the reduction the statistical error in the prediction of the protein domains, mostly in the case of the collection by Family.

Organisms

We analyse every row of the matrix and we create the histograms, using the definition in the section 3.3.3.

Families

To create the histograms, we conjecture that the organisms in a single family have the same parameter, so we define the Preston plot of a Family as the mean of the histograms of the organisms. First, we create the Preston plot for every organisms and we normalize the histogram with the value of the number of domains activated, developing the probability density function. A domain is activated when the appearance takes a non-zero value. Second, we addict the pdfs: the first column is the sum of all the first columns, the second is the sum of all second columns, and so on. Furthermore, the total area of the histogram has the value equal the number of organisms which stand in the Family. Therefore, the histogram is normalized by the number of organisms and now we can consider it as a probability density function, which is used in fitting. Finally, to aim a biological meaning of the ordinate, the final pdf is multiplied for the mean of the organisms' numbers of activated domains.

4.2.2 Method of cumulative distribution

We fit the Preston plot of our data set with the formula 3.43, paying attention that the first column contains $\langle\varphi_1\rangle$, the second contains $\langle\varphi_2\rangle + \langle\varphi_3\rangle$, and so on the ith column is given by

$$C_i = \sum_{2^{i-1}}^{2^i-1} \langle\varphi_i\rangle \tag{4.1}$$

4.2.3 Fitting in MATLAB

All the operations, described in the previous pages, are made by an algorithm, which we write in MATLAB language, after a manual exportation from a .tsv file, containing the data. For the fundamental operation of fitting, we used the function *fit()*, with the options shown below

```
FOption = fitoptions('Method', 'NonlinearLeastSquares',...
    'StartPoint', [0.5, 0.5], 'lower', [0,0], 'upper', [1,1]);
```

The first option is the method which the function search the best parameters. We decide tu use *Nonlinear Least Squares* method because of the construction of the function *RSAfit()*. The code of *RSAfit()* is avaible in the Appendix A. The default algorithm to resolve the minimization problem is the Levenberg-Marquardt one.

Finally, the class of options of the second row is formed by the initial condition and the bounds of the parameters. The bounds of the parameter $\frac{B}{D}$ are trivial, because of the model, on the other hand the upper limit of S is setted equal to 1 and there are not conflicts for our results. If a fit failed, probably the S value would take 1.

4.2.4 Uncertainty of parameters

Performing the function *fit()*, the 95% confidence of the parameters is returned by the function of MATLAB *confint()*.

4.2.5 Goodness of fits

To testing the goodness of the fits, we used the coefficient of determination R^2, evaluated as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \tag{4.2}$$

where y_i is the ith observed data, $\bar{y}$ is the mean of the observed data and $\hat{y}_i$ is the ith estimation.

Chapter 5

Results

5.1 Preston plot of Families

In this section are reported some of the histograms of the 115 Families.

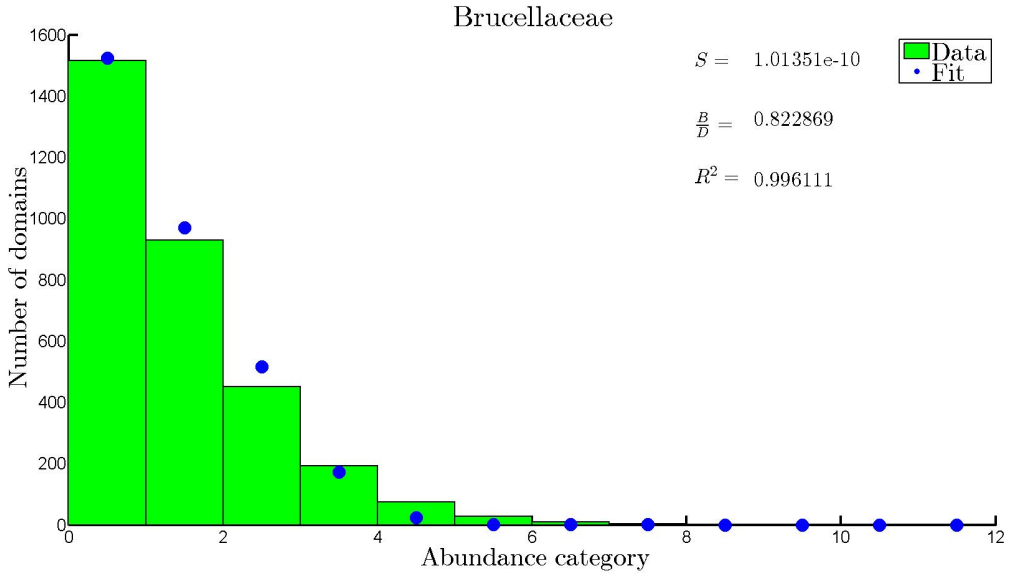

Figure 5.1: Preston plot of the Family Brucellaceae with the corresponding fit.

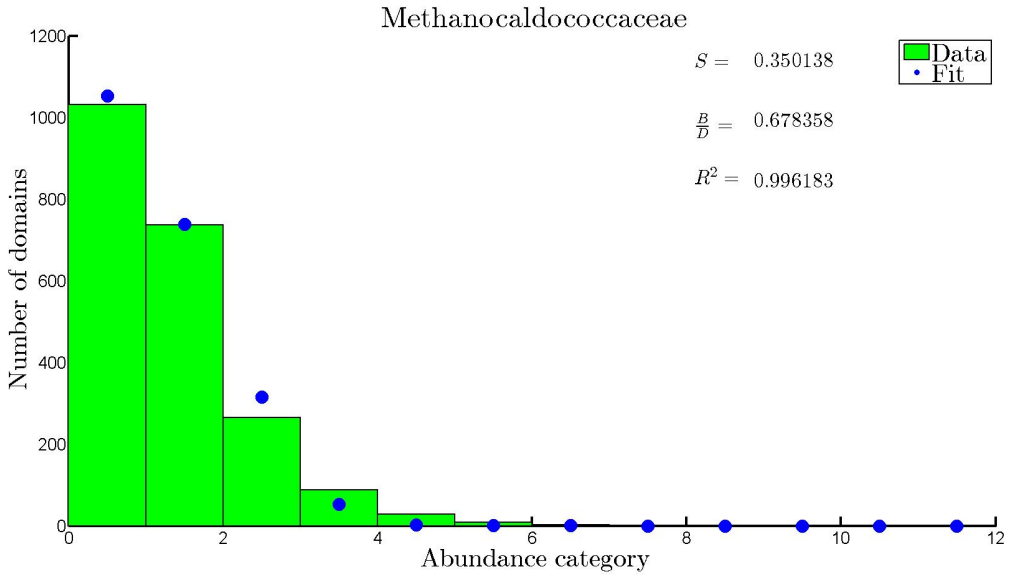

Figure 5.2: Preston plot of the Family Methanocaldococcaceae with the corresponding fit.

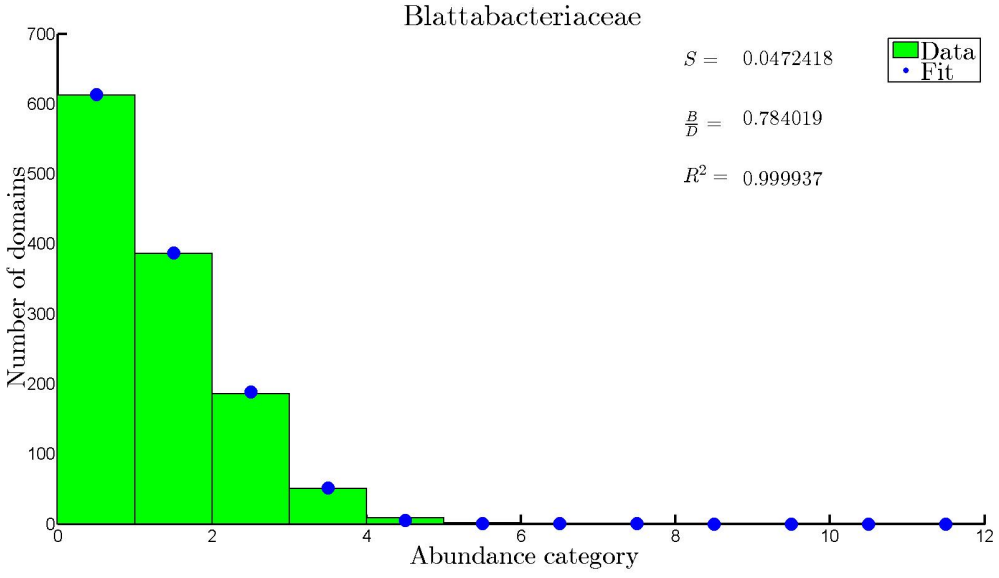

Figure 5.3: Preston plot of the Family Blattabacteriaceae with the corresponding fit.

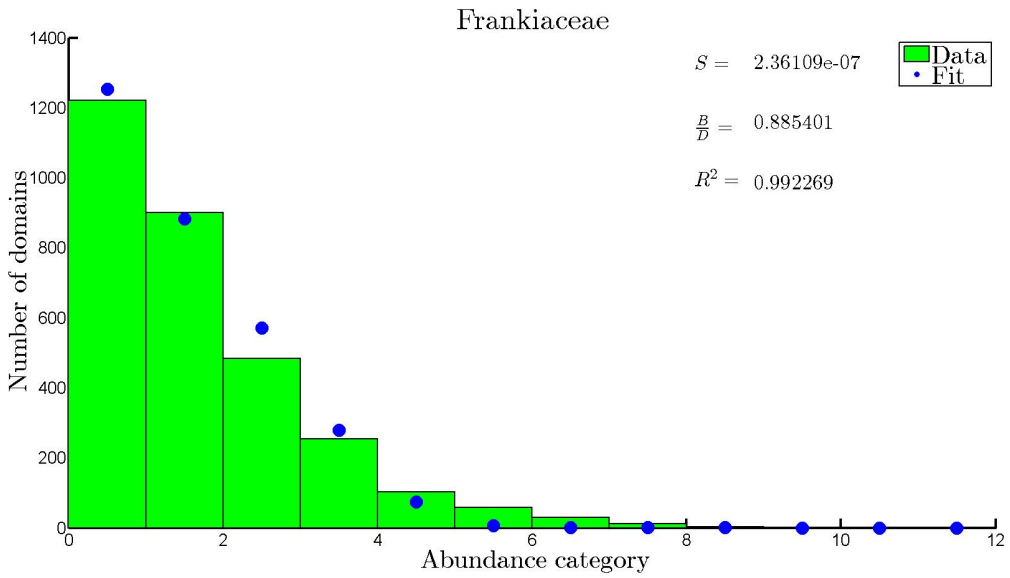

Figure 5.4: Preston plot of the Family Frankiaceae with the corresponding fit.

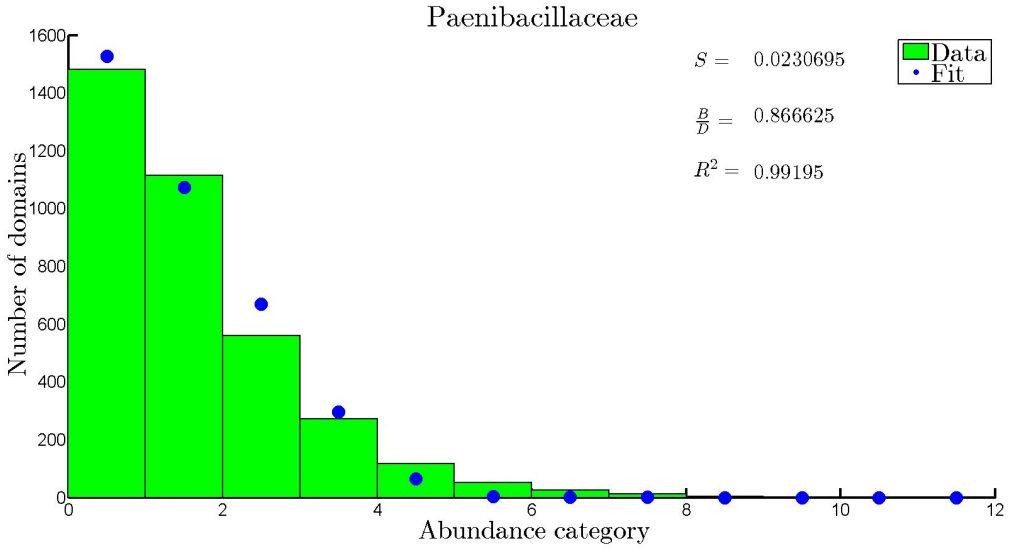

Figure 5.5: Preston plot of the Family Paenibacillaceae with the corresponding fit.

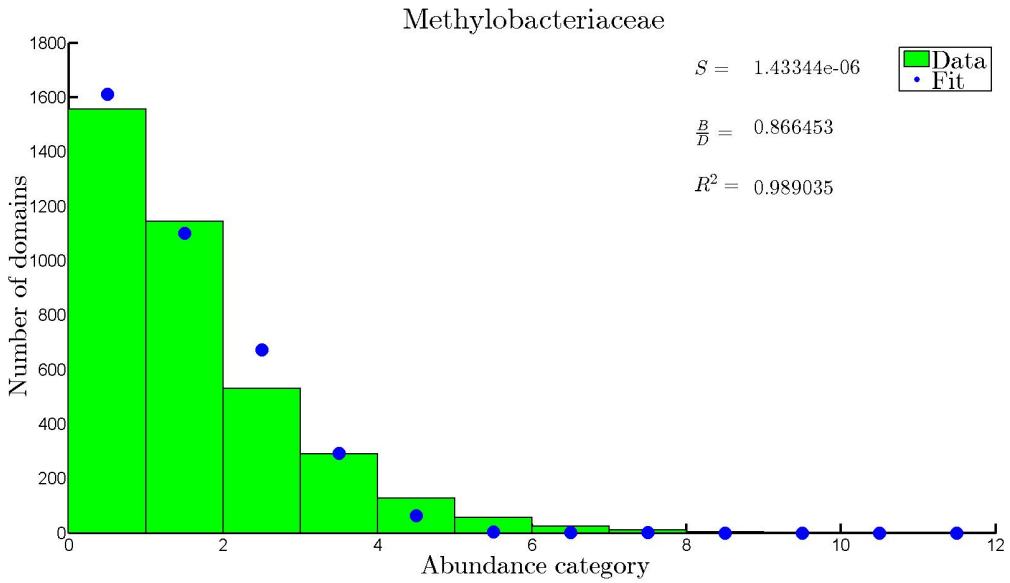

Figure 5.6: Preston plot of the Family Blattabacteriaceae with the corresponding fit.

5.2 Preston plot organism

In this section are reported some of the histograms of the more than 2300 organisms.

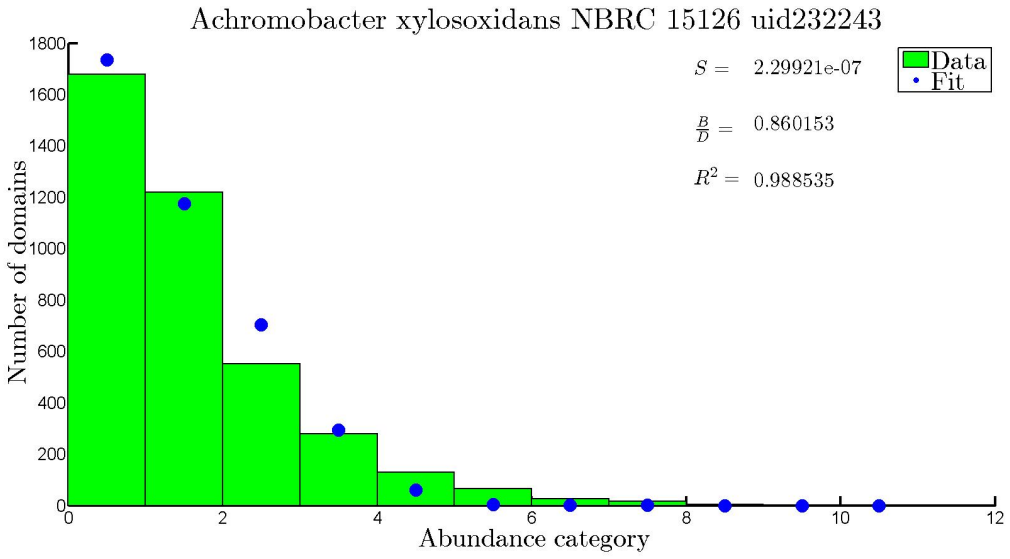

Figure 5.8: Preston plot of the organism Achromobacter xylooxidans with the corresponding fit.

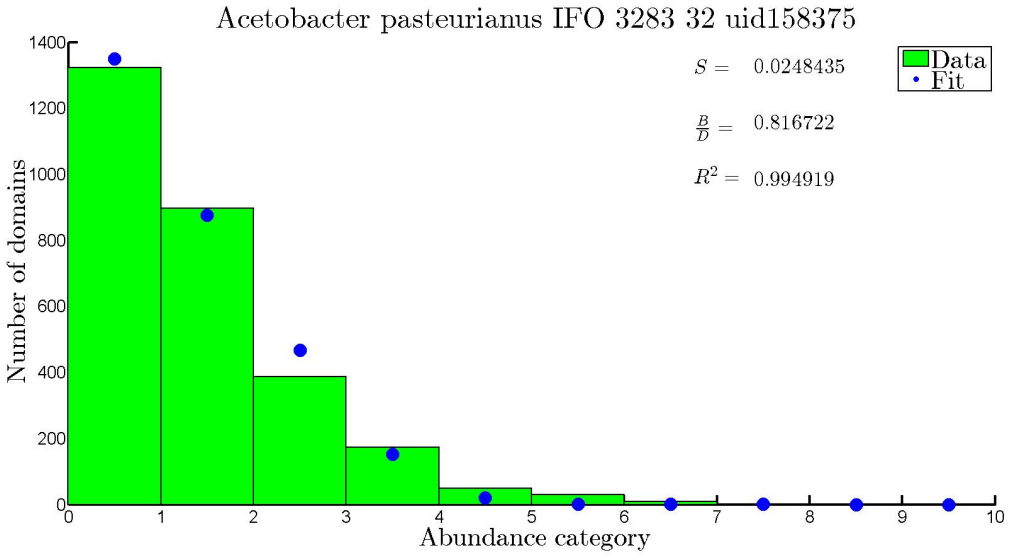

Figure 5.7: Preston plot of the organism Acetobacter pasteurianus with the corresponding fit.

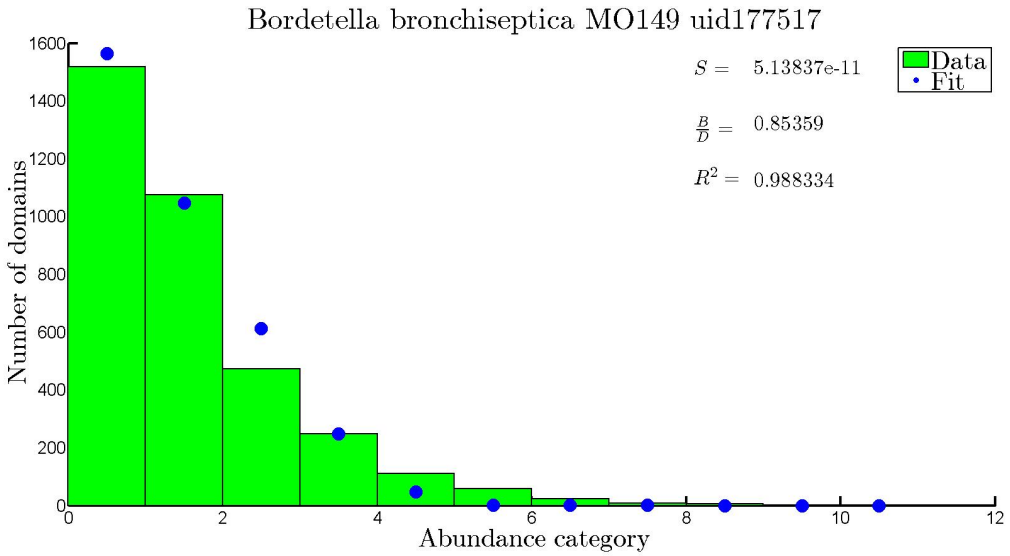

Figure 5.9: Preston plot of the organism Bordetella bronchiseptica with the corresponding fit.

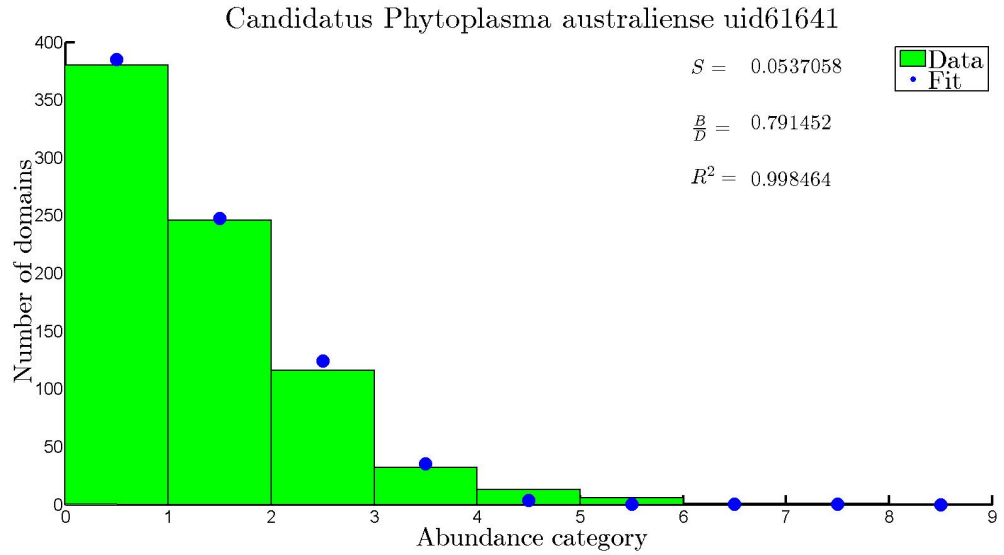

Figure 5.10: Preston plot of the organism Candidatus Phytoplasma with the corresponding fit.

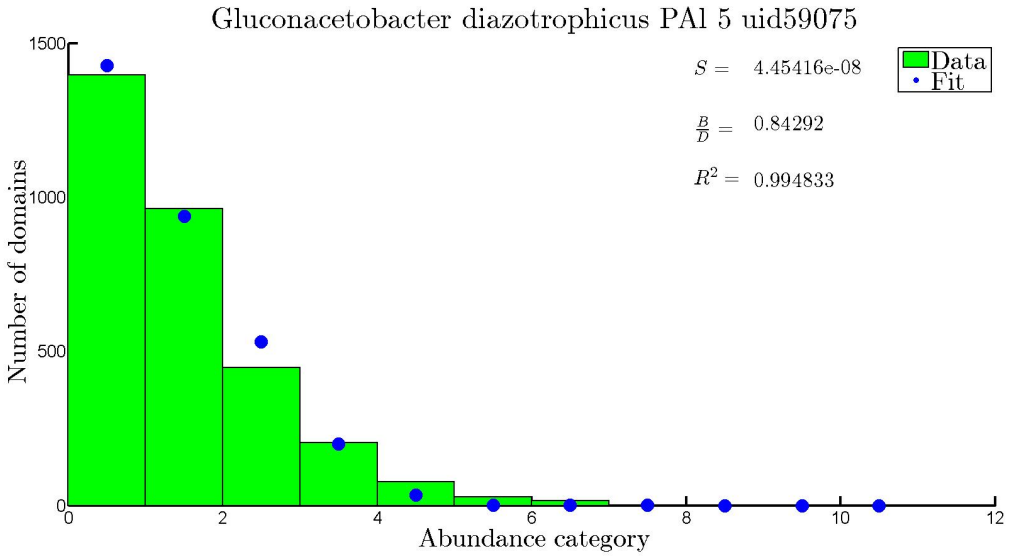

Figure 5.11: Preston plot of the organism Gluconacetobacter diazotrophicus with the corresponding fit.

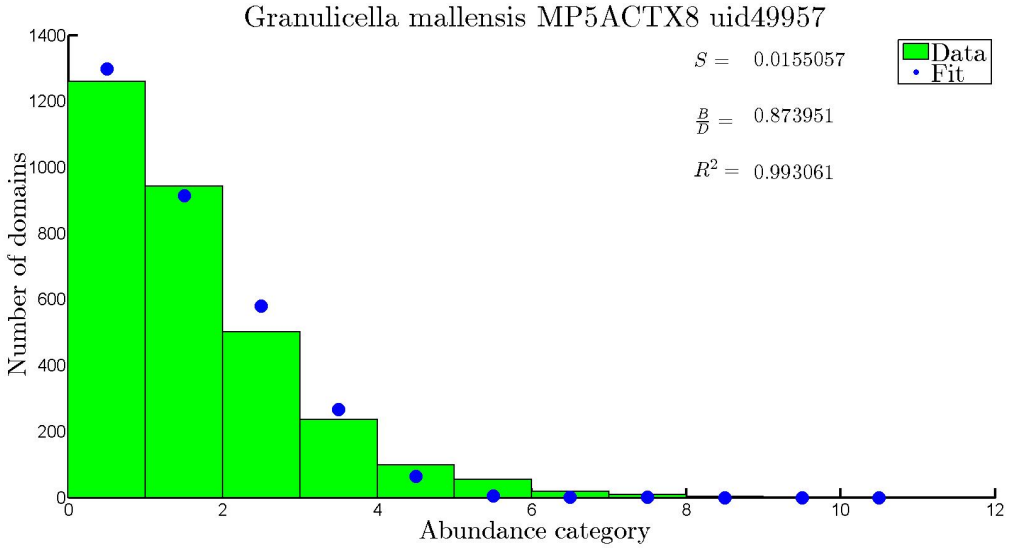

Figure 5.12: Preston plot of the organism Granulicella mallensis with the corresponding fit.

5.3 Families' parameters

In the Table 5.1 are shown the parameters' values computed by fitting, with their correspondent error and the value of R^2. The indetermination of the immigration parameter S for the Family *Brucellaceae* is not evaluated by the function, because the value of S is too small.

Family	S	ΔS	$\frac{B}{D}$	$\Delta\frac{B}{D}$	R^2
Acetobacteraceae	7,40E-07	0,13	0,832	0,050	0,9958
Acholeplasmataceae	0,110	0,10	0,777	0,039	0,9983
Acidithiobacillaceae	0,077	0,13	0,826	0,045	0,9961
Acidobacteriaceae	0,034	0,11	0,865	0,038	0,9953
Aeromonadaceae	0,031	0,15	0,831	0,057	0,9940
Alcaligenaceae	0,046	0,17	0,828	0,063	0,9927
Alteromonadaceae	0,028	0,17	0,834	0,062	0,9928
Anaplasmataceae	0,108	0,10	0,758	0,038	0,9987
Aquificaceae	0,081	0,085	0,805	0,032	0,9986
Archaeoglobaceae	0,185	0,17	0,757	0,059	0,9964
Bacillaceae	2,95E-04	0,14	0,854	0,052	0,9935
Bacteroidaceae	0,086	0,15	0,840	0,051	0,9936
Bartonellaceae	0,104	0,12	0,778	0,044	0,9979
Bifidobacteriaceae	1,39E-03	0,09	0,828	0,034	0,9982

Blattabacteriaceae	0,047	0,020	0,7840	0,0080	0,9999
Brachyspiraceae	1,12E-03	0,11	0,840	0,041	0,9968
Bradyrhizobiaceae	1,66E-03	0,15	0,856	0,054	0,9926
Brucellaceae	1,01E-10	NaN	0,823	0,013	0,9961
Burkholderiaceae	4,18E-04	0,15	0,866	0,054	0,9913
Campylobacteraceae	7,91E-04	0,051	0,818	0,021	0,9994
Caulobacteraceae	1,87E-03	0,14	0,844	0,052	0,9945
Chlamydiaceae	0,139	0,13	0,759	0,048	0,9978
Chlorobiaceae	0,105	0,12	0,797	0,042	0,9975
Chromatiaceae	0,100	0,15	0,814	0,053	0,9951
Clostridiaceae	0,123	0,15	0,823	0,051	0,9946
Comamonadaceae	0,011	0,15	0,849	0,055	0,9931
Coriobacteriaceae	0,068	0,11	0,821	0,042	0,9969
Corynebacteriaceae	0,023	0,097	0,829	0,037	0,9977
Coxiellaceae	0,170	0,17	0,765	0,061	0,9959
Cytophagaceae	6,89E-03	0,13	0,866	0,045	0,9939
Deferribacteraceae	0,053	0,089	0,827	0,033	0,9981
Dehalococcoidaceae	0,253	0,16	0,748	0,052	0,9970
Deinococcaceae	0,020	0,14	0,848	0,049	0,9944
Desulfobacteraceae	0,035	0,13	0,867	0,044	0,9936
Desulfobulbaceae	0,038	0,12	0,842	0,042	0,9960
Desulfovibrionaceae	7,62E-08	0,089	0,852	0,033	0,9975
Desulfurococcaceae	0,089	0,12	0,773	0,047	0,9978
Ectothiorhodospiraceae	0,101	0,14	0,797	0,052	0,9963
Enterobacteriaceae	0,029	0,18	0,825	0,068	0,9923
Enterococcaceae	0,104	0,18	0,812	0,065	0,9929
Eubacteriaceae	0,125	0,18	0,812	0,061	0,9932
Flavobacteriaceae	6,12E-04	0,089	0,844	0,034	0,9977
Francisellaceae	0,192	0,15	0,764	0,052	0,9970
Frankiaceae	2,36E-07	0,13	0,885	0,042	0,9923
Fusobacteriaceae	0,165	0,12	0,782	0,043	0,9975
Geobacteraceae	0,023	0,11	0,861	0,040	0,9954
Halobacteriaceae	0,080	0,13	0,839	0,045	0,9953
Halomonadaceae	9,19E-03	0,048	0,755	0,021	0,9997
Helicobacteraceae	0,055	0,082	0,804	0,031	0,9987
Hyphomicrobiaceae	5,16E-07	0,13	0,842	0,050	0,9952
Lachnospiraceae	0,133	0,17	0,819	0,058	0,9931
Lactobacillaceae	0,109	0,16	0,809	0,056	0,9947
Legionellaceae	0,056	0,16	0,824	0,058	0,9941
Leptospiraceae	3,79E-07	0,099	0,850	0,037	0,9970
Leuconostocaceae	0,105	0,13	0,799	0,049	0,9966

Listeriaceae	0,080	0,17	0,817	0,060	0,9938
Methanobacteriaceae	0,231	0,18	0,735	0,061	0,9966
Methanocaldococcaceae	0,350	0,24	0,678	0,076	0,9962
Methanococcaceae	0,236	0,19	0,718	0,066	0,9967
Methanosarcinaceae	0,134	0,15	0,802	0,052	0,9956
Methylobacteriaceae	1,43E-06	0,17	0,866	0,060	0,9890
Methylophilaceae	0,138	0,17	0,791	0,059	0,9951
Microbacteriaceae	4,13E-07	0,10	0,845	0,039	0,9969
Micrococcaceae	6,97E-04	0,11	0,859	0,040	0,9958
Micromonosporaceae	4,48E-07	0,13	0,891	0,041	0,9918
Moraxellaceae	0,080	0,18	0,813	0,065	0,9931
Mycobacteriaceae	1,25E-06	0,13	0,867	0,047	0,9932
Mycoplasmataceae	0,038	0,060	0,778	0,024	0,9995
Myxococcaceae	0,017	0,14	0,877	0,047	0,9914
Neisseriaceae	0,208	0,16	0,745	0,057	0,9969
Nitrosomonadaceae	0,085	0,13	0,806	0,049	0,9964
Nitrospiraceae	0,040	0,090	0,833	0,033	0,9979
Nocardiaceae	5,00E-07	0,13	0,884	0,045	0,9916
Nostocaceae	0,049	0,16	0,858	0,056	0,9906
Oceanospirillaceae	1,03E-06	0,14	0,843	0,053	0,9945
Oxalobacteraceae	0,075	0,14	0,809	0,053	0,9958
Paenibacillaceae	0,023	0,15	0,867	0,050	0,9919
Pasteurellaceae	0,221	0,17	0,745	0,060	0,9964
Peptococcaceae	0,097	0,14	0,834	0,047	0,9949
Peptostreptococcaceae	0,091	0,15	0,825	0,053	0,9943
Phyllobacteriaceae	7,50E-07	0,14	0,867	0,050	0,9924
Piscirickettsiaceae	0,143	0,17	0,778	0,060	0,9956
Planctomycetaceae	2,60E-04	0,11	0,872	0,040	0,9948
Porphyromonadaceae	0,102	0,12	0,806	0,043	0,9971
Prevotellaceae	0,072	0,14	0,812	0,051	0,9959
Prochlorococcaceae	0,071	0,11	0,783	0,042	0,9982
Propionibacteriaceae	7,52E-03	0,089	0,843	0,034	0,9977
Pseudomonadaceae	0,035	0,18	0,840	0,065	0,9909
Pseudonocardiaceae	5,05E-06	0,13	0,894	0,042	0,9910
Rhizobiaceae	8,31E-07	0,13	0,855	0,048	0,9943
Rhodobacteraceae	4,63E-07	0,15	0,852	0,054	0,9932
Rhodocyclaceae	0,051	0,15	0,840	0,054	0,9935
Rhodospirillaceae	0,013	0,14	0,850	0,053	0,9935
Rhodothermaceae	5,20E-04	0,083	0,855	0,031	0,9977
Rickettsiaceae	8,68E-04	0,089	0,779	0,038	0,9988
Ruminococcaceae	0,160	0,17	0,800	0,060	0,9940

Shewanellaceae	0,060	0,19	0,821	0,071	0,9913
Sphingobacteriaceae	0,048	0,13	0,857	0,044	0,9945
Sphingomonadaceae	4,00E-03	0,11	0,840	0,043	0,9965
Spirochaetaceae	6,47E-03	0,064	0,839	0,024	0,9989
Spiroplasmataceae	0,084	0,088	0,775	0,034	0,9989
Staphylococcaceae	0,105	0,15	0,805	0,055	0,9954
Streptococcaceae	0,025	0,12	0,818	0,047	0,9968
Streptomycetaceae	1,09E-03	0,13	0,902	0,041	0,9896
Sulfolobaceae	0,155	0,18	0,788	0,062	0,9944
Synergistaceae	0,113	0,11	0,792	0,041	0,9978
Thermaceae	0,063	0,12	0,828	0,043	0,9965
Thermoanaerobacteraceae	0,137	0,13	0,797	0,047	0,9966
Thermoanaerobacterales	0,079	0,11	0,812	0,039	0,9976
Thermococcaceae	0,201	0,24	0,743	0,083	0,9935
Thermoproteaceae	0,131	0,12	0,792	0,045	0,9972
Thermotogaceae	0,057	0,092	0,815	0,035	0,9982
Veillonellaceae	0,205	0,16	0,773	0,054	0,9961
Vibrionaceae	0,071	0,19	0,818	0,069	0,9920
Xanthomonadaceae	9,10E-03	0,14	0,837	0,053	0,9947

Table 5.1: Best fitting parameters for every Family

We plot the space of parameters and we notice a linear correlation, that it is displayed in the Figure 5.13.

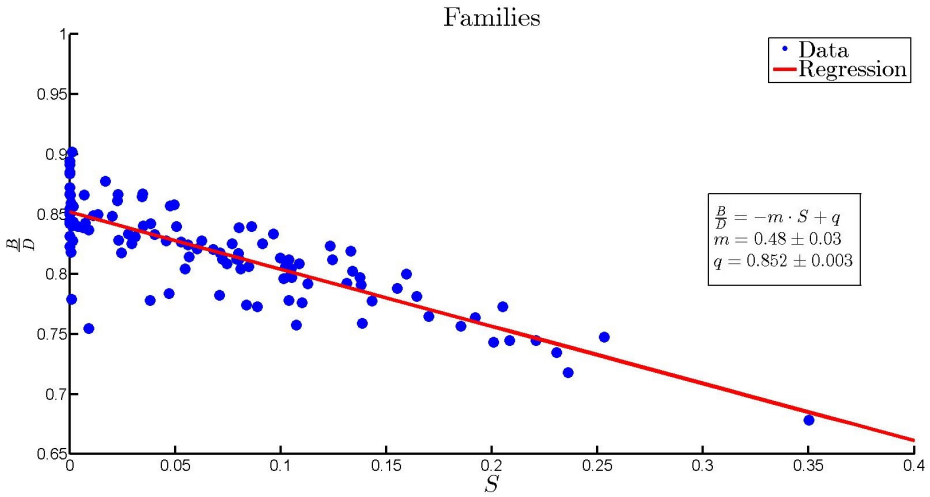

Figure 5.13: Scatter plot of the Families' parameters. The red curve is the linear regression.

5.3.1 R^2

In the Figure 5.14 the values of R^2 are ordered from the smaller to the biggest. The minimum value is greater than 0.988.

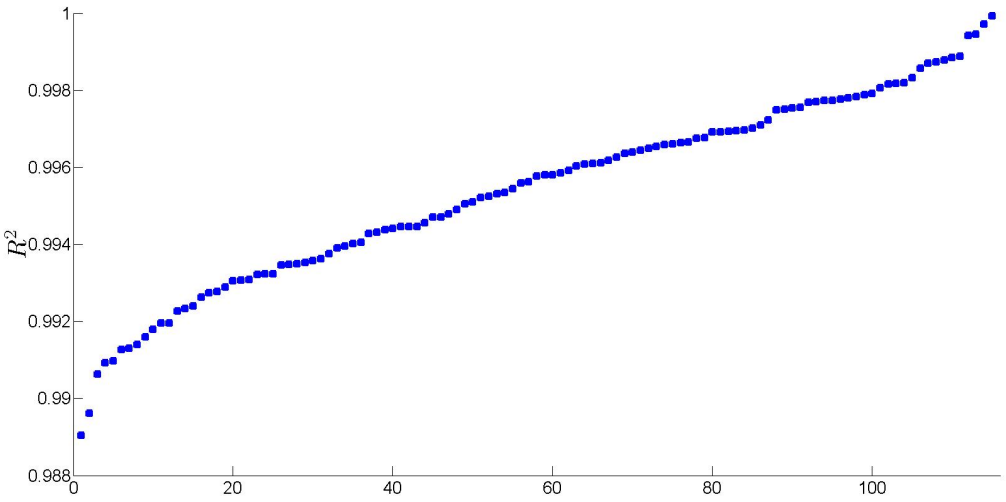

Figure 5.14: Run ordered plot of the values of R^2 for the Families.

5.4 Organisms' parameters

Because of the great number of organisms (more than 2300) there is not a table where the parameters are reported: the table would occupy more than forty pages. The Figure 5.15 shows how the parameters are distributed. The points with the same color belong to the same Family.

5.4.1 R^2

In the Figure 5.16 the values of R^2 are ordered from the smaller to the biggest. The minimum value is greater than 0.982.

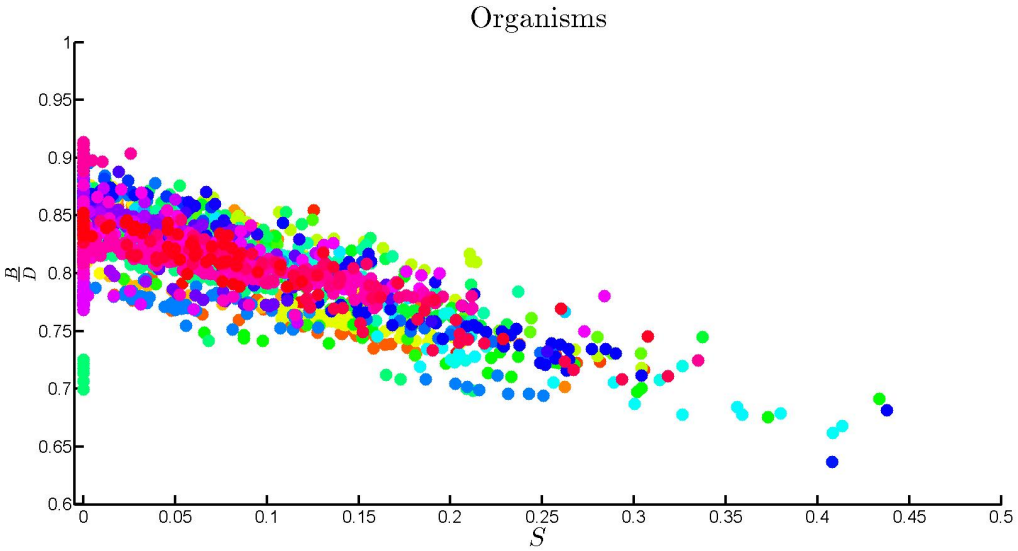

Figure 5.15: Scatter plot of the organisms' parameters. However there are very similar shades, the organisms of the same Family have the same color in the plot.

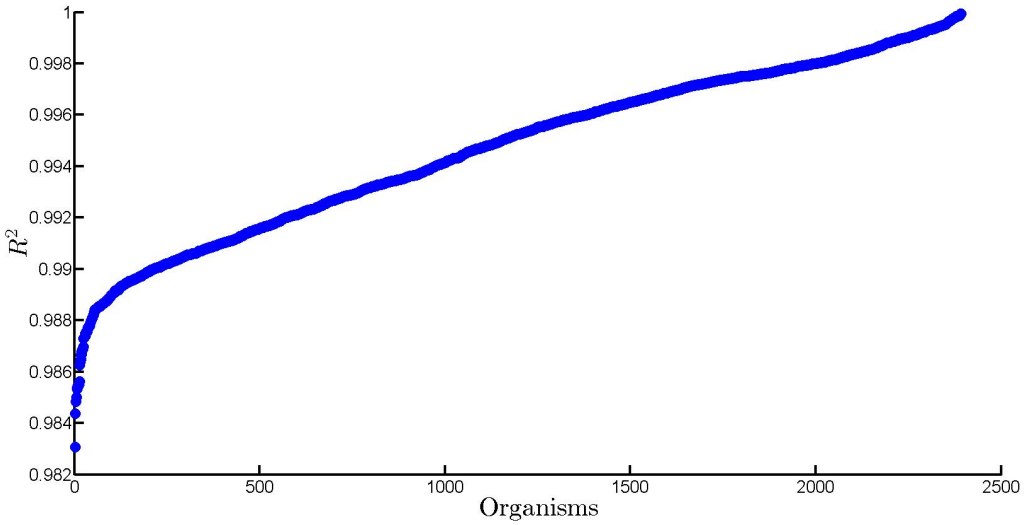

Figure 5.16: Run ordered plot of the values of R^2 for the organisms.

5.5 Null model

The creation of a null model lets us test our hypothesis. Starting from the data, the null model is formed by different shufflings of the columns of the matrix of the domains appearances. This type of recombination means that the organisms exchange their protein domains in a random way and they loose information about the Family which belong. In other word, we attempt a total randomly switching of genetic material and grouping the organisms without a logical pattern, certainly without a phyletic meaning. Afterwards, we replicate every steps used for the "real" data, illustrated in the section 4.2.

The result in the parameters' space (Figure 5.17) is without any doubt a good results: the Families seem as if they have the same values of parameters, except the statistic fluctuation.

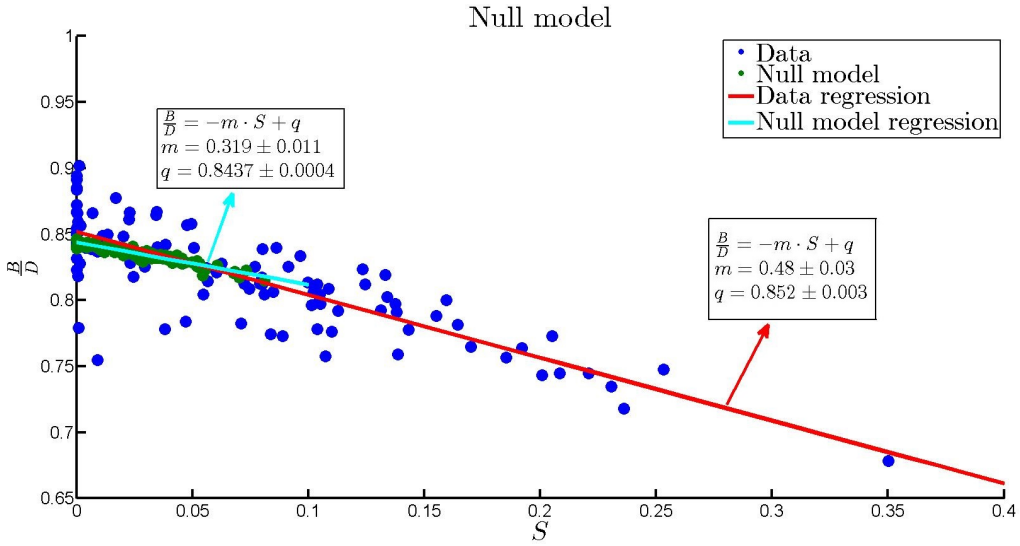

Figure 5.17: The green points are the parameters fitted from the null model's data. They occupy a very little part of the entire space than the parameters obtained from real data (in blue).

We think the best result is the comparison between the distribution of the parameters, which are shown in the Figure 5.18. The real data distribution reminds a power law distribution, while the distribution of the null model's parameter S looks like a Gaussian distribution, that confirm a fluctuation trend around the mean value.

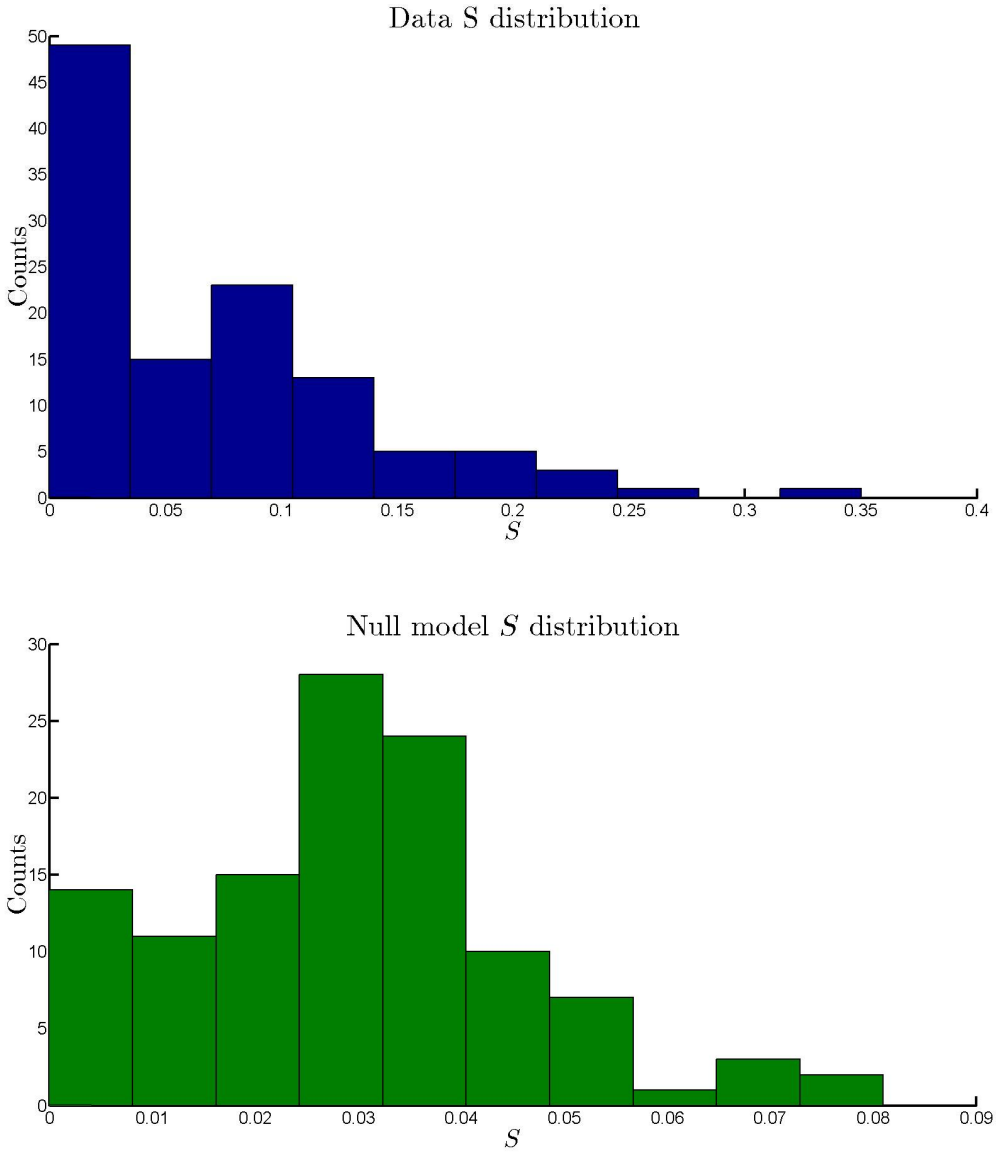

Figure 5.18: The up graph is the distribution of the immigration parameter evaluated from real data and the down one is the distribution of the same parameter in the null model.

Chapter 6

Conclusions

We work on a data set which has never been used before. So we characterize for the for first time the protein domains, in an ecological point of view.

The principal column of our work is the ecological neutral theory of the creation and the disappearance between the domains during the evolution, in an entire genome. Moreover this theory is suggestive and is gaining ground, day by day, within the international scientific community.

We assume that the system (genome) is in a steady state in this time scale, where the events that guide the changes in the population have to be seen from an evolutionary time scale. Each bacterial genome, as protein domains' distribution, represents a system in steady state and any other steady state would be classified as a different strain.

The model we proposed has various advantages, such as its simplicity. We use a linear model in the description of the birth and death rates, adding a migration parameter. The main obstacle was the normalization condition: we had to evaluate the total number of possible domains as function of something we had (the number of observed domains).

Furthermore we obtain great results in fitting the histograms. For every Preston plot, the R^2 values are very close to 1 (the unrealistic number of perfection), both for the Families' ones and for the organisms' ones.

On the other hand, we think that the best result is the comparison between our fits and the null model's ones. The totally random exchange would differentiate fewer the Families, creating a Gaussian fitness of the distribution of the migration parameter S, instead of a power law distribution. This behaviour implies that what we find is not something accidental, but is closely bound with the intrinsic characteristics of the data set analysed.

We suppose that the relation between the ratio of birth end death rates and the migration parameter has a powerful biological meaning. Indeed, the Families characterized by a few numeric value of the migration parameter, during the evolution, probably, have overcome this disadvantage with a greater value of the birth ratio. On the contrary, the Families characterized by a few numeric value of the birth ratio, presumably, have adjust

themselves showing a greater value of the migration parameter.

We think that the value of the migration parameter has two different sense, one biological and the other computational. First, we suppose that this parameter is tightly related with the rate of horizontal gene transfer: greater is the numeric value, greater is the percentage of the genes transferred horizontally. In literature we find that in the Families with the greatest values the presence of horizontal gene transfer with Archea is confirmed: Nesbø et al. [39] for *Dehalococcoidaceae*, Wolf et al. [53] for *Methanobacteriaceae*, *Methanocaldococcaceae*, *Methanococcaceae* and *Thermococcaceae*, Hotopp et al. [25] for *Neisseriaceae*.

Second, the distributions with migration parameter close to zero suggest us an computational aim. A zero value of the migration parameters means that the number of the total number of domains, which potentially may be in the genome, tends towards infinity ($\propto (cost)^S$, with $0 < cost < 1$). According with the fact that usually half of the proteins in the genome are hypothetical proteins and have no domain assigned, we suppose that in the Families with the lowest value of the migration parameter there are still many domains to be identified and annotated. So, this number could give a hint on which families we should look in more detail and sequence more.

Finally, we think that keeping on studying this data base with this approach could help evolutionary researches and, maybe, could do the groundwork for a new taxonomy, based on the horizontal gene transfer.

Appendix A

RSAfit.m

Below is reported the code of the function using for fitting our data. The high values of k reached from data require prefering the logarithmic value of function Γ (MATLAB function *gammaln()*) instead of the factorial, using the relation $n! = \Gamma(n+1)$.

```
function Phi = RSAfit (N, Y, X)

Lunghezza = length (N);

Phi = zeros (Lunghezza,1);

for i = 1: 1 : Lunghezza
    Inizio = 2^(N(i)-1);
    Fine = 2^N(i) -1;
    for k = Inizio: 1 : Fine
        Phi(i) = Phi(i) + (X^k)*
                 *exp(gammaln(k+Y)-gammaln(k+1))/(((1-X)^(-Y)-1)*gamma(Y));
    end
end

end
```

Bibliography

[1] L. J. S. Allen. *An introduction to stochastic processes with applications to biology*. Pearson Education New Jersey, 2003.

[2] P. Artymiuk, A. R. Poirrette, D. W. Rice, and P. Willett. A polymerase i palm in adenylyl cyclase? *Nature*, 388:33–34, 1997.

[3] S. Azaele, S. Pigolotti, J. R. Banavar, and A. Maritan. Dynamical evolution of ecosystems. *Nature*, 444:926–928, 2014.

[4] N. T. J. Bailey. *The elements of stochastic processes with applications to the natural sciences*, volume 25. John Wiley & Sons, 1990.

[5] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 1966.

[6] C. Boschetti, A. Carr, A. Crisp, I. Eyres, Y. Wang-Koh, E. Lubzens, T. G. Barraclough, G. Micklem, and A. Tunnacliffe. Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genetics*, 8, 2012.

[7] C. I. Branden and J. Tooze. *Introduction to Protein Structure*. Garland, 1999.

[8] L. L. Cavalli-Sforza. The dna revolution in population genetics. *Trends in Genetics*, 14(2):60–65, 1998.

[9] J. M. Chase and M. A. Leibold. *Ecological niches: linking classical and contemporary approaches*. University of Chicago Press, 2003.

[10] C. Chothia and M. Levitt. Structural patterns in globular proteins. *Nature*, 261:552–558, 1976.

[11] C. Chothia, M. Levitt, and D. Richardson. Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences*, 74(10):4130–4134, 1977.

[12] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[13] A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, and G. Micklem. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(1):50, 2015. Alastair Crisp and Chiara Boschetti contributed equally to this work.

[14] T. Dagan, Y. Artzy-Randrup, and W. Martin. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences*, 105(29):10039–10044, 2008.

[15] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing discrete-time finite markov chains. *Journal of Applied Probability*, 2(1):pp. 88–100, 1965.

[16] J. Davies. Origins and evolution of antibiotic resistance. *Microbiologia (Madrid, Spain)*, 12(1):9—16, March 1996.

[17] J. Davison. Genetic exchange between bacteria in the environment. *Plasmid*, 42(2):73–91, 1999.

[18] S. Domingues, K. M. Nielsen, and G. J. da Silva. Various pathways leading to the acquisition of antibiotic resistance by natural transformation. *Mobile Genetic Elements*, 2(6):257–260, 2012.

[19] J. C. Dunning Hotopp. Horizontal gene transfer between bacteria and animals. *Trends in Genetics*, 27:157–163, 2015.

[20] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13429, 1996.

[21] W. Feller. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1):11–40, 1939.

[22] 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. *Journal of Animal Ecology*, 12(1):42–58, 1943.

[23] R. Grantham, C. Gautier, and M. Gouy. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8:1893–1912, 2010.

[24] C. Gyles and P. Boerlin. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology Online*, 51(2):328–340, 2014.

[25] J. C. D. Hotopp, R. Grifantini, N. Kumar, Y. L. Tzeng, D. Fouts, E. Frigimelica, M. Draghi, M. M. Giuliani, R. Rappuoli, D. S. Stephens, G. Grandi, and H. Tettelin. Comparative genomics of neisseria meningitidis: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology*, 152(12):3733–3749, 2006.

[26] S. P. Hubbell. *The unified neutral theory of biodiversity and biogeography (MPB-32)*, volume 32. Princeton University Press, 2001.

[27] D. Hyatt, G. Chen, P. LoCascio, M. Land, F. Larimer, and L. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.

[28] T. Itoh, K. Takemoto, H. Mori, and T. Gojobori. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution*, 16(3):332–346, 1999.

[29] P. Jones, D. Binns, H. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Yong, R. Lopez, and S. Hunter. Interproscan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9):1236—1240, 2014.

[30] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin. Birth and death of protein domains: A simple model of evolution explains power law behavio. *BMC Evolutionary Biology*, 2:18–18, 2002.

[31] D. Kidane, S. Ayora, J. Sweasy, P. L. Graumann, and J. C. Alonso. The cell pole: The site of cross talk between the dna uptake and genetic recombination machinery. *Critical reviews in biochemistry and molecular biology*, 47:531–555, 2012.

[32] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology*, 55(1):709–742, 2001. PMID: 11544372.

[33] I. Letunic and P. Bork. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, 39:W475–W478, 2011.

[34] H. Maughan and R. J. Redfield. Tracing the evolution of competence in *Haemophilus influenzae*. *PLoS ONE*, 4(6):e5854, 2009.

[35] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71, 1 1958.

[36] D. Moreira and H. Philippe. Molecular phylogeny: pitfalls and progress. *International Microbiology*, 3(1):9–16, 2010.

[37] J. Mrázek and S. Karlin. Detecting alien genes in bacterial genomesa. *Annals of the New York Academy of Sciences*, 870(1):314–329, 1999.

[38] I. Nasell. Extinction and quasi-stationarity in the verhulst logistic model. *Journal of Theoretical Biology*, 211(1):11–27, 2001.

[39] C. L. Nesbo, S. L'Haridon, K. O. Stetter, and W. F. Doolittle. Phylogenetic analyses of two "archaeal" genes in thermotoga maritima reveal multiple transfers between archaea and bacteria. *Molecular Biology and Evolution*, 18(3):362–375, 2001.

[40] A. S. Novozhilov, G. P. Karev, and E. V. Koonin. Biological applications of the theory of birth-and-death processes. *Briefings in Bioinformatics*, 7(1):70–85, 2006.

[41] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[42] C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, 31(1):45–71, 2002. PMID: 11988462.

[43] F. W. Preston. The commonness, and rarity, of species. *Ecology*, 29(3):254–283, 1948.

[44] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Research*, 33(suppl 2):W116–W120, 2005.

[45] C. Sala, G. Castellani, and D. Remondini. Ecological modelling for next generation sequencing data, 2013.

[46] B. F. Smets and T. Barkay. Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. *Nat Rev Micro*, 3:675–678, 2005.

[47] S. Smith. The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. *The FASEB Journal*, 8(15):1248–59, 1994.

[48] M. J. Stanhope, A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411:940–944, 2001.

[49] M. Syvanen. Horizontal gene transfer: Evidence and possible consequences. *Annual Review of Genetics*, 28(1):237–261, 1994. PMID: 7893125.

[50] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[51] S. Venner, C. Feschotte, and C. Biémont. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics*, 25(7):317–323, 2009.

[52] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan. Patterns of relative species abundance in rainforests and coral reefs. *Nature*, 450:45–49, 2007.

[53] Y. I. Wolf, K. S. Makarova, N. Yutin, and E. V. Koonin. Updated clusters of orthologous genes for archaea: a complex ancestor of the archaea and the byways of horizontal gene transfer. *Biology direct*, 7, 2012.

[54] Y. I. Wolf, I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, 11(3):356–372, 2001.

[55] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.