Gene families distributions across bacterial genomes : from models to evolutionary genomics data

Eleonora de Lazzari

Université Pierre et Marie Curie

Ecole doctorale Physique en Île-de-France (ED564)

Laboratoire de Biologie Computationnelle et Quantitative (UMR 7238)

Gene families distributions across bacterial genomes: from models to evolutionary genomics data.

Thèse de doctorat de Physique

Par: Eleonora De Lazzari

Présentée et soutenue publiquement le: 9 Novembre 2017

Devant un jury composé de:

Amos MARITAN	*Examinateur*
Namiko MITARAI	*Examinateur*
Marco COSENTINO LAGOMARSINO	*Directeur de thèse*
Aleksandra WALCZAK	*Rapporteur*
Bianca SCLAVI	*Rapporteur*
Didier CHATENAY	*Rapporteur*

Abstract

Comparative genomics has established itself as a fundamental discipline to unravel evolutionary biology. The first challenge to overcome a mere descriptive knowledge of comparative genomics data is to develop a higher-level description of the content of a genome. For this purpose we employed the modular representation of genomes to explore quantitative laws that regulate how genomes are built from elementary functional and evolutionary ingredients.

The first part of the work sets off from the observation that the number of domains sharing the same functional annotation of a genome increases as a power law of the genome size. Since functional categories are aggregates of domain families, we asked how the abundance of domains performing a specific function emerges from evolutionary moves at the family level. We found that domain families are also characterized by family-dependent scaling laws, supporting the idea that genome evolution occurred under the interplay of constraints over functional and evolutionary families.

The following chapter aims to provide a general theoretical framework for the emergence of shared components from dependency in empirical component systems with non-binary abundances. In order to do this, we defined a positive model that builds a realization from a set of components linked in a dependency network. The ensemble of resulting realizations reproduces both the distribution of shared components and the law for the growth of the number of distinct families with genome size.

Finally, the last chapter attempts to extend the component systems approach to microbial ecosystems, i.e., sets of genomes sharing the same environments. Using our findings about domain families scaling laws, we analyzed how the abundance of domain families in a metagenome is affected by the constraint of power-law scaling of family abundance in individual genomes. The result is the definition of an observable, whose functional form contains access quantitative information on the original composition of the metagenome.

Contents

Chapter 1

Introduction

The first complete bacterial genome, that of *Haemophilus influenzae Rd*, was sequenced in 1995 [21] followed by the genome of *Mycoplasma genitalium* in the same year [24]. Subsequently, the collection of sequenced genomes rapidly progressed and reached in 1999 a steady exponential growth [43]. This massive information was essential to the development of comparative genomics since it allowed to identify sets of of orthologs (genes that evolved from the same ancestral gene) and to determine which gene families are present or absent in a particular genome [44]. The comparative analysis of genomes has opened new perspectives in evolutionary biology, to the point that it has been defined as "the only route to satisfactory reconstructions of evolution" [43].

However, the abundant information from the many available genome sequences may be very difficult to understand. The first step to overcome a mere descriptive knowledge is to develop a higher-level description of the content of a genome. Using gene families as a choice for the constitutive building blocks has led to the notable finding of several simple quantitative laws [43]. These empirical laws allows to get some insight into the "recipes" by which genomes are built from elementary functional and evolutionary ingredients. Specifically, it has been bound that the number of families associated with the same biological function scales as a power law of the genome size, calculated as the total count of domains [82]. Depending on the function examined, the scaling exponent varies from 0 to 2. The power-law distribution characterizes also other genomic quantities such as the total number of families found on a genome [16] and the distribution of family sizes [36, 45, 16], as well as the distribution of family occurrence, i.e. the fraction of genomes sharing a given fraction of families [65].

Hence, the exploitation of such modular representations of genomes has opened new perspectives for the understanding of genome evolution. The explanation of these empirical regularities requires models that mimic basic steps of genome evolution, such as gene loss and duplication[45, 16, 30, 65]. From the viewpoint of statistical physics, there is an opportunity to explore new models

whose scope is to capture the salient ingredients of the empirical laws as null or positive features (and thus improve our understanding of genome composition.)

The same set of studies suggest that a possible general framework [16, 30] is to represent genomes as "component systems" i.e. modular architectures considered as sets of elementary units. Here, we will take this approach, and we will consider some statistical features of the set of proteins found on all sequenced genomes, in terms of structural protein domains [63]. Protein domains are the modular sub-shapes forming the building blocks for proteins. They constitute independent thermodynamically stable structures, and are not expected to split over the course of evolution [75], because of their physical stability and function. Hence, domains may be used as convenient "evolutionary atoms" [45], because they typically do not split into smaller units or form by fusing multiple copies of other domains [10], as proteins can do. A domain also determines a set of potential biochemical or functions and interactions for a protein, such as binding and participation in well-defined classes of chemical reactions [63].

Combining information about shapes, functions and sequences of protein domains, it is possible to generate a systematic hierarchical classification of protein domains [2, 19, 62]. Generally, this classification comprises three layers. At the lowest level, domains are grouped into families on the basis of significant sequence similarity. Families with lower sequence identity but whose structures and functional features suggest a common evolutionary origin, are grouped into superfamilies. Finally, domains belonging to superfamilies or families are defined as having a common fold if they share the same major secondary structures. The repertoire of basic secondary structures for domains seems to be relatively small [45].

In brief, protein domains are a convenient coarse-grained representation of proteins, which contains information on their evolution and function. Describing proteins in terms of domains is very useful, and has become almost commonplace in biology. The "component system" approach extrapolates this coarse-grained representation to the description of whole genomes and sets of genome. The modular structure shown in Fig. 1.1 will be the core of this thesis.

This work is divided into three main chapters, where different questions are asked and where the same data structure is considered from different viewpoints. The first one (Chapter 2) sets off from the observation that the number of domains sharing the same functional annotation of a genome increases as a power law of the genome size [57, 71, 82]. The scaling exponents are function-dependent and are directly linked with the probability that the addition/deletion of a domain would be fixed over evolutionary times. Since functional categories are aggregates of domain families, we asked how the abundance of domains performing a specific function emerges from evolutionary moves at the family level. We found that domain families are also characterized by family-dependent scaling laws, supporting the idea that genome evolution occurred under the interplay of constraints over functional and evolutionary families.

The goal of Chapter 3 is to provide a general theoretical rationale for the emergence of shared components from dependency in empirical component systems with redundancy, i.e., where components can appear more than once in realizations, as is the case for protein domains in genomes. In order to develop this theoretical framework, we extended the ideas of Pang and Maslov [65] to the case of components with non-binary abundance, and defined a positive model that builds a

realization from a set of components linked in a dependency network. The ensemble of resulting realizations reproduces both the distribution of shared components and the law for the growth of the number of distinct families with genome size.

Finally, Chapter 4 attempts to extend the component systems approach to microbial ecosystems, i.e., sets of genomes sharing the same environments. Using our findings about domain families scaling laws (described in Chapter 2), we analyzed how the abundance of domain families in a metagenome is affected by the constraint of power-law scaling of family abundance in individual genomes. The result is the definition of an observable, whose functional form contains access quantitative information on the original composition of the metagenome, specifically the moments of the distribution of genome sizes.

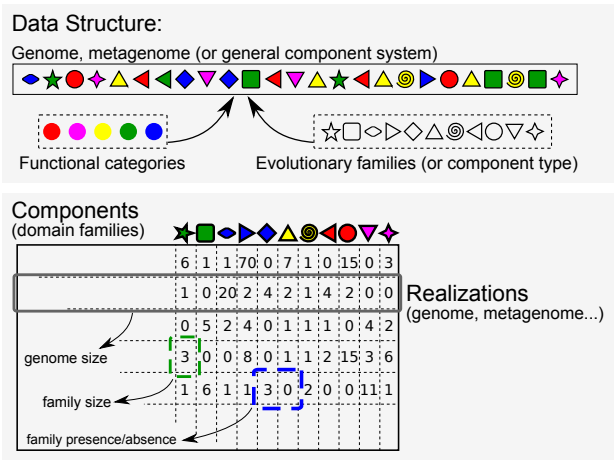

Figure 1.1: Modular systems. Representation of a system into components and realizations. Top: example of a realization. Shapes represent components (for example, the gene families on a genome) and they can be found multiple times in the realization (the genome in our example), defining the abundance of a given component. Colors represent possible further functional annotations of components. Bottom: component abundance matrix, where several realization (for example several sequenced genomes from several species) can be compared based on the abundance and presence/absence of theoretical modular components.

Chapter 2

Family-specific scaling laws in bacterial genomes.

2.1 Introduction

2.1.1 High-level functional categories of genes follows quantitative laws

As demonstrated by van Nimwegen [82] and confirmed by a series of follow-up studies [57, 58, 13, 30, 11], striking quantitative laws exist for high-level functional categories of genes. Specifically, the number of genes within individual functional categories exhibit clear power-laws, when plotted as a function of genome size measured in terms of its number of protein-coding genes or, at a finer level of resolution, of their constitutive domains (see Fig. 2.1).

In prokaryotes, such scaling laws appear well conserved across clades and lifestyles [58], supporting the simple hypothesis that these scaling laws are universally shared by this group. From the evolutionary genomics viewpoint [42], these laws have been explained as a byproduct of specific

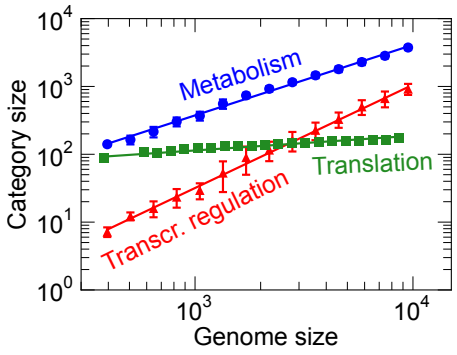

Figure 2.1: Functional category scaling laws This plot shows the mean number of protein-domains associated with functional categories (y-axis) "translation" (green squares), "metabolism" (blue circles), and "transcription regulation" (red triangles) as a function of the total number of domains in the genome (x-axis). Both axes are shown on a logarithmic scale. The straight lines show power-law fits. Each functional category has a well defined power law scaling with function specific exponent: regulation of transcription scales quadratically, metabolic domains increase linearly with the genome size and translation remains constant.

"evolutionary potentials", i.e., per-category-member rates of additions/deletions fixed in the population over evolution. As predicted by quantitative arguments, estimates of such rates correlate well with the category scaling exponents [82, 57].

A complementary point of view [55, 64, 30] focuses on the existence of universal "recipes" determining ratios of proteins between different functions necessary for genome functionality . Such recipes should mirror the "dependency structure" or network operating within genomes as well as other complex systems [65]. According to this point of view the usefulness, and thus the occurrence, of a given functional component depends on the presence of a set of other components, which are necessary for it to be operational.

2.1.2 The analysis of quantitative laws at the domain-family level may explain how the scaling of functional categories emerges from the evolutionary dynamics.

Beyond functional categories, protein coding genes can be classified in "evolutionary families" defined by the homology of their sequences. Functional categories usually contain genes from tens or more of distinct evolutionary families.

The statistics of gene families also exhibits quantitative laws and regularities starting from a universal distribution of their per-genome abundance [36], explained by evolutionary models accounting for birth, death, and expansion of individual families [69, 38, 16]. While some earlier work connects per-genome abundance statistics of families with functional scaling laws [30], the link between functional category scaling and evolutionary expansion of gene families that build them remains relatively unexplored. Clearly, selective pressure is driven by functional constraints, and thus selection cannot in principle recognize families with identical functional roles. On the other hand, slight differences in the functional spectrum of different protein domains, and interdependency of different functions can make the scenario more complex. Thus, one central question is how the abundance of genes performing a specific function emerges from the evolutionary dynamics at the family level. Two alternative extreme scenarios can be put forward:

(*i*) The high-level scaling laws could emerge only at the level of functions, and be "combinatorially neutral" at the level of the evolutionary families building up a particular function. In this case all or most of the families performing a particular function would be mutually interchangeable.

(*ii*) Functional categories scaling could be the result of the sum family-specific scaling laws. Therefore the evolutionary potentials would be family-specific and coincide with family evolutionary expansion rates, possibly emerging from the complex dependency structure cited above, and from fine-tuned functional specificity of distinct families.

An intermediate possibility is that an interplay of constraints acts on both functional and evolutionary families. The following sections address the question of which is the most likely scenario by providing a systematic analysis of scaling laws at the family level and their interplay with functional category scaling. We will focus only on bacteria.

2.2 Families have individual scaling exponents, reflected by family-specific scaling laws

We started by addressing the question of whether individual families show scaling laws, and thus can be associated to specific scaling exponents. In this section we will first present the methods and then discuss the first result, that is the existence of family scaling laws.

2.2.1 Data analysis

symbol	category code	category name	symbol	category code	category name
☆	C	Energy	▽	OA	Proteases
✚	E	Amino acids met./tr.	▬	P	Ion met./tr.
▲	F	Nucleotide met./tr.	⬡	RA	Redox
◀	G	Carbohydrate met./tr.	⬮	RB	transferase
✖	H	Coenzyme met./tr.	■	RC	Other enzymes
⬬	J	Translation	♦	RF	Transport
◆	L	DNA replication	⬟	S	Unknown function
▼	LA	DNA binding	●	T	Signal transduction
▮	O	Protein modification			

Table 2.1: Symbols and codes used to identify functional categories.

We considered bacterial proteomes retrieved from the SUPERFAMILY (release 1.75 downloaded in October 2014, [27]) and PFAM (release 27.0 downloaded in October 2014, [7, 19]) database. Evolutionary families were defined from the domain assignments of 1535 superfamilies (SUPERFAMILY database) and 446 clans (PFAM database) on all protein sequences in completed genomes. We focused the analysis on the 1112 bacterial proteomes used as species reference in the SUPERFAMILY database. For the functional annotations of the SUPERFAMILY data, we considered annotation of SCOP domains as a scheme of 50 more detailed functional categories, mapped to 7 more general function categories, developed by C. Vogel [84]. Functional categories will be usually identified by a one- or two-letter code that we retrieved from [84]. In Table 2.1 we listed the functional category descriptive name, its code and the symbol associated. PFAM clans were annotated on the same scheme of 50 functional categories, using the mapping of clans into superfamilies available from the PFAM website `http://pfam.xfam.org/clan/browse#numbers` [20].

For each evolutionary domain family (or a functional category consisting of multiple evolutionary families), genome sizes (measured in the overall number of domains) were logarithmically binned. For each bin we calculated mean and standard deviation of the given family abundance (number of domains) within the bin. The estimated scaling exponent β_i for family i is the result

of the non-linear least squares fitting of the binned data weighted by the standard error of family abundance. Genome size bins containing less than 10 genomes were not taken into account.

Many domain families are found only in a few genomes and/or in very few copies. For this reason, they might not show clear scaling properties. We excluded such families from the analysis with some filtering criteria. In order to filter out these families, we used three independent parameters.

(*i*) *Occurrence.* The occurrence of the family i is defined as:

$$o_i = \frac{N_G^{(i)}}{N_G} \,, \tag{2.1}$$

while $N_G^{(i)}$ is the number of genomes in which family i is present and N_G is the total number of genomes in the sample.

(*ii*) *Correlation.* We defined ρ_i as the Pearson correlation coefficient ρ_i between the logarithm of the family abundance and the logarithm of the genome size.

(*iii*) *Goodness of fit.* First let's define the quantity LS_i as

$$LS_i = \frac{1}{N_G^{(i)}} \sum_{g=1}^{N_G^{(i)}} \frac{\left[y_{fit,i} - y_i\right]^2}{y_{fit}} \,,$$

where y_i is the logarithm of the empirical abundance of family i and $y_{fit,i}$ is the abundance calculated with the fit parameters, i.e.,

$$y_{fit,i} = A_i + \beta_i \log\left(\sum_{i=1}^{N_F} n_i^g\right) .$$

The goodness of fit index s_i was defined as

$$s_i = \frac{1}{1 + \sqrt{LS_i}}$$

so that the values of s_i close to 1 correspond to the minimum value of the average squared deviation between the fit and the empirical values.

It should be noted that the correlation ρ_i and the goodness of fit s_i are independent from o_i as shown in Fig. 2.2A-B. The filter over the occurrence assures that each family is present in the majority of the genomes and thus that there are enough points to infer a scaling behavior. The correlation ρ_i quantifies the existence of a relation between family abundance and genome size. However, considering families with clear but shallow scaling or constant abundance across genomes, ρ_i gives values close to zero, or slightly negative, therefore another parameter is required to assess the accuracy of the fit results. The index s_i puts on the same ground families with different exponents,

but generally decreases as the scaling exponent increases, in accordance with the growth of fluctuations in families with higher exponents observed in ref. [31]. Hence, we decided to use it only for low exponents, where the Pearson correlation is a bad proxy of scaling.

We considered only the families with $o_i > 0.6$. If the fitted scaling exponent is higher than 0.2 then we excluded the families with $\rho_i < 0.4$, while, for exponents lower than 0.2, only families with $s_i > 0.9$ were taken into account. After this thresholding, we removed 1179 families out of 1536. While the fraction of such small and sparse families is large, we verified that they do not contribute significantly to the category scaling (see Figure 2.2C).

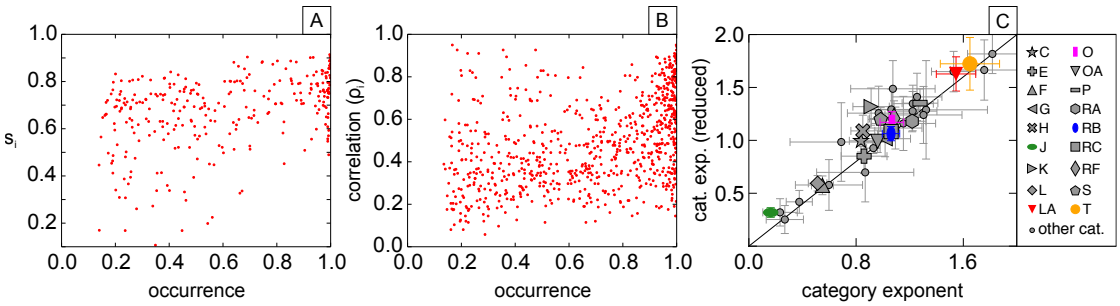

Figure 2.2: The parameters used to filter out families are not correlated and the removal of filtered families does not influence the functional category scaling.(A) The plot reports the goodness of fit index s_i, which is the average squared deviation between the empirical family abundance and the one derived from the fit, as a function of family occurrence. Each point represents a family whose exponent is lower than 0.2 . (B) Pearson correlation between family abundance (number of domain belonging to a given family) and genome size, calculated across the genomes where the family is present as a function of family occurrence. Each point represents a family whose exponent is higher than 0.2 . The lack of clear correlation visible in the plots shows that the three indices are all relevant in the filters. (C) The plot compares the category exponent obtained by considering all the domains and the exponent obtained by removing from the category count the domains belonging to families filtered out by our criteria for unclear scaling. The exponents before (x-axis) and after thresholding (y-axis) are compatible within their errors. The solid line is the $y = x$ line. The panel on the right shows the association between symbols and category codes (see Table 2.1 for the corresponding category name).

2.2.2 Comparison with a null model supports the existence of scaling laws at the family level is not simply due to sampling effects.

The families that pass the quality filtering procedure all show a clearly identifiable individual scaling when plotted as a function of genome size. As an example, Fig. 2.3 shows the scaling of a set of chosen families in four selected functional categories. It is worth noting that some low-abundance families that occur in all genomes with a very consistent number of copies show definite scaling with exponents close to zero [31], being clearly constant with size, with little or no fluctuations.

Additionally, Fig. 2.3 shows that the presence of "outlier families" is common among functional categories. In most categories, we found families where the deviations from the category exponents is clear, beyond the uncertainty due to the errors from the fits. Fig. 2.3 shows some

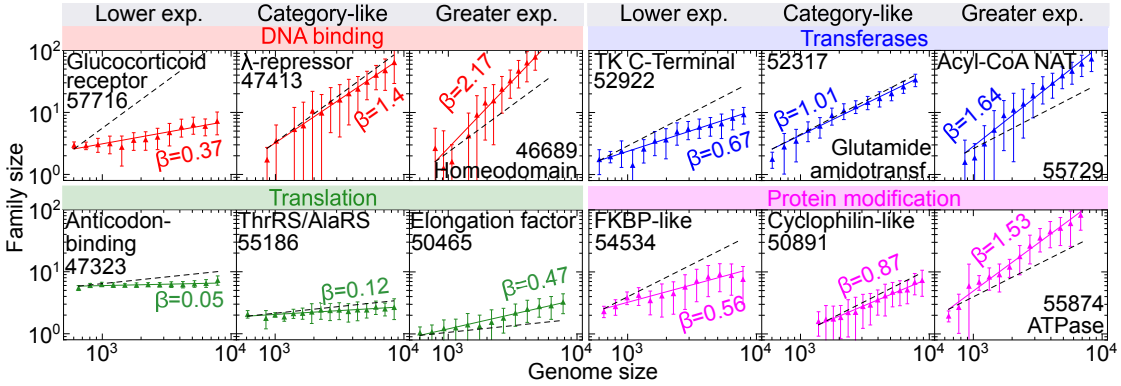

Figure 2.3: Families follow specific scaling laws, which may agree or deviate from the overall scaling of the functional category to which the family belongs. The plots report the abundance of twelve different superfamilies as a function of the genome size (triangles are binned averages). The power-law fits (colored lines) are compared to the power-law fits of the functional category to which each family belong (dashed black lines). We display here examples from four functional categories: DNA binding (top row), Translation (second row from top), Transferases (third row from top) and Protein modification (bottom row). Families in the leftmost / rightmost column scale respectively slower/faster than their category means, families in the middle column have similar slope to the full category. Legends specify the SCOP superfamily id, family descriptive name and power-law exponent (β_i) from the fits. The lines for the functional categories were shifted in order to intersect the family scaling law at its minimum x value. The original intercepts are: 0.0007 ± 0.0143 (DNA binding), 39.131 ± 0.006 (Translation), 0.04 ± 0.02 (Transferases) and 0.01 ± 0.03 (Protein modification).

examples where in each of the shown categories β_i is higher, lower or comparable to β_c. A table containing all the family and category exponents is available at appendix A.

Given that functional categories follow specific scaling laws, likely related to function-specific evolutionary trends [82, 57], there remain different open possibilities for the behavior of the evolutionary families composing the functional categories. One simple scenario is that family scalings are family-specific, thus validating the existence of family evolutionary expansion rates that are quantitatively different to the one of their functional category. In the opposite extreme scenario the scaling is only function-specific, and individual families performing similar functions are interchangeable. If this were the case, the observed family diversity in scaling exponent would be only due to sampling effects. To assess the influence of sampling effects, we defined a null model, in which we randomized the families within a category conserving their occurrence patterns and the category average abundance. In more detail, the null model is based on the following ingredients:

(*i*) The number of domains belonging to a category c in genome g, n_i^g, is conserved.

(*ii*) For each genome, domains are not assigned to families that are not present in that genome.

(*iii*) The average frequency $f_c(i)$ for each family i with respect to the category c is conserved.

This quantity is defined as:

$$f_c(i) = \frac{1}{N_G^{(i)}} \sum_g \frac{n_i^g}{n_c^g} \,, \tag{2.2}$$

where the family index i belongs to the set in category c and the sum over g is carried over all the genomes, while $N_G^{(i)}$ is the number of genomes in which family i is present, n_i^g is the abundance of family i in genome g and $n_c^g = \sum_{i \in c} n_i^g$ is the abundance of category c in genome g.

Note that the occurrence and the average frequency are uncorrelated in the data, hence we chose to conserve both. Given a genome g, each realization of the null model redistributes randomly the n_c^g domains of the functional category c arranged in the F_c^g families belonging to category c in genome g. Each one of the n_c^g domains is assigned to family i with probability

$$p_c(i) = \begin{cases} \dfrac{f_c(i)}{\sum_{i \in c} f_c(i)}, & \text{if } n_i^g \neq 0 \\ 0, & \text{if } n_i^g = 0 \,. \end{cases} \tag{2.3}$$

The randomized families always show very similar scaling as the one of the corresponding category (see Fig. 2.4). Hence, this analysis strongly supports the existence of family-specific scaling exponents that do not simply descend from the category scaling.

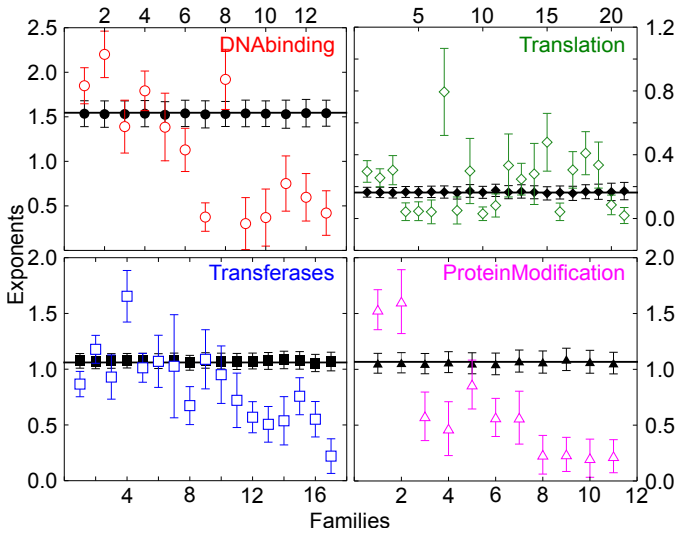

Figure 2.4: Family exponents differ significantly from the null expectation set by the scaling of the associated functional category. In order to account for random fluctuations in family composition within a category, the family-exponents (empty symbols with error bars) were compared with the ones calculated randomizing the data accordingly to the null model presented in this section (black squares, error bars are variability across 1000 realizations). The variability obtained from the null model is extremely low and is not sufficient to explain the variability of scaling exponents of different families within a category. Each panel corresponds to a different functional category, its scaling exponent is shown as the black horizontal line. Families within each category are sorted in decreasing order of abundance, i.e. total domain count in the category.

2.2.3 Family exponents correlate with diversity of biochemical functions but not with contact order or evolutionary rate of domains.

Finally, we considered the correlation of family scaling exponents with relevant biological and evolutionary parameters. We tested the diversity of EC-numbers associated with families, quantifying the functional plasticity of a given family. The Enzyme Commission (EC) number is a classification scheme for enzyme-catalyzed chemical reactions. It is built as a four-levels tree where the top nodes are six main groups of reactions, namely Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases [6]. We used the mapping between Superfamilies and EC terms [27], to investigate the correlation between the Superfamily scaling and the number of different reactions in which the family is involved. This quantity is the count of distinct EC numbers corresponding to the finest level of the EC classification. Table 2.2 shows the correlations with other parameters such as foldability (quantified by size-corrected contact order, SMCO [17]), selective pressure (quantified by the ratio of nonsynonymous to synonymous K_a/K_s substitution rates [61]) and overall family abundance.

The results are summarized in Table 2.2. Foldability and K_a/K_s appears to have little correlation with scaling exponents. Instead, we found a significant positive correlation of exponents with family abundance, and both quantities are correlated with diversity of EC-numbers in metabolic families. This suggests that, at least for metabolism, functional properties of a fold play a role in family scaling, and that beyond metabolism, abundance and scaling are, on average, not unrelated.

database	*parameters*	β_i	A_i	$\langle a_i \rangle$	f_i/o_i
SUPFAM	SMCO	−0.06	0.07	0.04	0.04
	EC numbers diversity (not met. families)	0.22	−0.14	0.40	0.35
	EC numbers diversity (met. families)	0.64	−0.50	0.77	0.74
PFAM	Hmm length	0.13	−0.13	0.10	0.10
	Ka/Ks	0.11	−0.12	0.03	0.05

Table 2.2: Spearman correlations among family parameters. The table reports Spearman correlation coefficients between sets of family parameters, comparing biological/evolutionary and abundance properties. Each row describes biological parameters: for the Superfamily database we used the foldability and the diversity of EC-numbers associated with families. For Pfam families, we considered the Hidden Markov Model sequence length (Hmm length) and the evolutionary rate K_a/K_s. The parameters listed in columns are the exponent and prefactor of the family scaling law (β_i and A_i respectively), the mean family abundance calculated over all genomes ($\langle a_i \rangle$) and the ratio between the average relative abundance (see definition of frequency in Section 2.2.2) and family occurrence (f_i/o_i). Relevant correlations are found for the diversity in EC numbers restricted to metabolic families and the scaling exponent β_i, as well as with the mean and relative family abundance. Family abundance and scaling exponent are also correlated (Spearman 0.72).

2.3 The heterogeneity in scaling exponents is function-specific.

The analyses presented in the previous section support the hypothesis that functional categories contain families with specific scaling exponents. Indeed, the scaling exponents β_i of the families can be significantly different from the category exponent β_c, with deviations that are much larger than predicted by randomizing the categories according to the null model (see Fig. 2.4).

In order to quantify this "scaling heterogeneity" of functional categories, we computed for each family i the distance between its scaling exponent β_i and the category exponent β_c:

$$h_i = |\beta_c - \beta_i|,$$

Finally, we defined an index H_c quantifying the heterogeneity of the scaling of the families within a category by averaging this distance over the families associated to a given category c:

$$H_c = \frac{1}{F_c} \sum_i h_i,$$

where F_c is the number of families in category c.

Figure 2.5A shows the relation between the heterogeneity H_c and the category exponent β_c. Interestingly, these two quantities are correlated, with categories with larger values of β_c being more heterogeneous. Intuitively, categories with small exponents are incompatible with extremely large fluctuations of family exponents, while categories with larger exponents can contain families with small β_i. Indeed, this trend of heterogeneity with exponents is also observed in the null model, where the heterogeneity of null categories is much smaller than empirical ones, since all families tend to take the exponent category (Fig. 2.4).

Figure 2.5B allows a direct comparison of the heterogeneity of different categories by subtracting the mean trend. It is noteworthy that the Signal Transduction functional category, which also has clear superlinear scaling, has much lower heterogeneity than DNA-binding/transcription factors. Among the categories with linear scaling, Transferases is one of the least heterogeneous ones, while the categories Protein Modification and Ion metabolism and Transport show a large variability in the exponents of the associated families. For Protein Modification, this signal is essentially due to the Gro-ES superfamily and to the HFSP90 ATP-ase domain, which have a clear superlinear scaling, while other chaperone families, such as FKBP, HSP20-like and J-domain are clearly sublinear with exponents close to zero. Interestingly, the Gro-EL domains, functionally associated to the Gro-EL, are part of this second class (exponent close to 0.2), showing very different abundance scaling to the Gro-EL partner domains. Conversely, the category Ion Metabolism and Transport is divided equally into linearly scaling (e.g., Ferritin-like Iron homeostasis domains) and markedly sublinear families, such as SUF (sulphur assimilation) / NIF (nitrogen fixation) domains. On the other hand, categories with small values of heterogeneity are made of families with exponents close to the one of the category, as shown in Table 2.3 in the case of, e.g., Transferases.

We also observe that, since the total abundance of a category is the sum of the abundances of the corresponding families, one may see a conceptual inconsistency in stating that both families

and categories are well described by power laws. Indeed, a sum of power-laws can only be a power law if all the exponents are identical. On the other hand, these constraints describe a mean behavior.

The inclusion of the fluctuation terms around the means makes the two scalings not formally inconsistent. Additionally, the narrower the distribution of family exponents within a category, the better the power-law approximation should hold at the functional categories level. Consequently, we tested the connection of category heterogeneity in exponents to goodness of fit. We found that the Spearman correlation coefficient between category heterogeneity and mean residual of the fit is equal to 0.43, indicating that more heterogeneous categories give slightly worse fits as expected by these considerations.

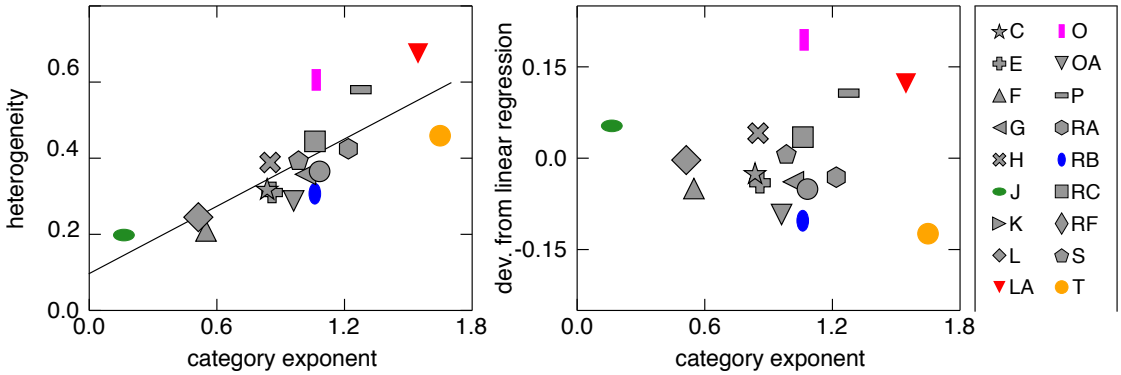

Figure 2.5: (A) Functional categories with faster scaling laws contain families with more heterogeneous scaling exponents. Heterogeneity is quantified by the mean deviation between the family scaling exponents and the category exponent. The plot reports heterogeneity scores for different functional categories, plotted as a function of the category exponents. The black line is the linear fit between heterogeneity and exponents (slope 0.3, intercept 0.1). (B) Comparison of heterogeneities subtracted from the linear trend. By this comparison, the least heterogeneous categories are Signal Transduction (T) and Transferase (RB), and the most heterogenous are DNA Binding (LA) and protein modification (O). Translation (J) is slightly above the trend for its low exponent. The legend (right panel) shows the association between symbols and category codes (see Table 2.1 for the corresponding category name).

2.4 Determinants of the scaling exponent of a functional category

We have shown that scaling exponents of individual families may correspond to a variable extent to the exponent of the corresponding functional category. However, since categories are groups of families, the scaling of the former cannot be independent of the scaling of the latter. This section explores systematically the connection between the two. As detailed below, we find that in some cases the scaling exponent of functional categories is determined by few outlier families, while in other cases most of the families within a category contribute to the category scaling exponent.

While many families have a clear power-law scaling, functional categories may contain many low-abundance families with unclear scaling properties. When considered individually, these families do not contribute much to the total number of domains of a category, but their joined effect on the scaling of the category could be potentially important. Fig. 2.7 shows that the sum of these low-abundance families does not suffer from sampling problems and shows a clear scaling. Interestingly, the scaling exponents for these sums once again does not necessarily coincide with the category exponents.

Figure 2.6A illustrates the systematic procedure that we used in order to understand how the scaling of categories emerges from the scaling of the associated families. Families were ranked by total abundance across all genomes (from the most to the least abundant) and removed one by one from the category. At each removal step in this procedure, both the scaling exponent of the removed family and the exponent of the remainder of the category are considered. In other words, the i-th step evaluates the exponent of the i-th ranking family (in order of overall abundance) and of the set of families obtained by removing the i top-ranking families (with highest abundance) from the category. The resulting exponents quantify the contribution of each family to the global category scaling, as well as the collective contribution of all the families with increasingly lower overall abundance.

The results (Fig. 2.6B), show how the heterogeneity features described above are related to family abundance. Pooled together, the low-abundance families within a functional category may show very different scaling than their category. Additionally, single families follow scaling laws that deviate from the one of the corresponding functional categories. One notable example of this are Transcription-Factor DNA-binding domains. If the abundance of the outliers families is large enough in terms of the fraction of domains in the functional category, they might be responsible for determining the scaling of the entire category, as it happens in the case of DNA-binding (which is more extensively discussed in the following section).

Overall, one can distinguish between two main behaviors, either a category scaling is driven by a low number of highly populated "outlier" families (e.g. DNA binding and Protein Modification in Fig. 2.6B), or the category scaling is coherent, and robust to family subtraction (e.g. Transferases and Translation in Fig. 2.6B). While the first behavior appears to be more common for functional categories with higher scaling exponent, there are some exceptions. Notably, the scaling of strongly super-linear categories is not always driven by a few families. For example, the functional category Signal Transduction has an exponent $\beta_c = 1.7$, which remains stable after the removal of the largest families (Fig. 2.7). Both behaviors are clearly visible for intermediate exponents (in order

to appreciate this, compare the Transferases and Protein Modification categories in Fig. 2.6B).

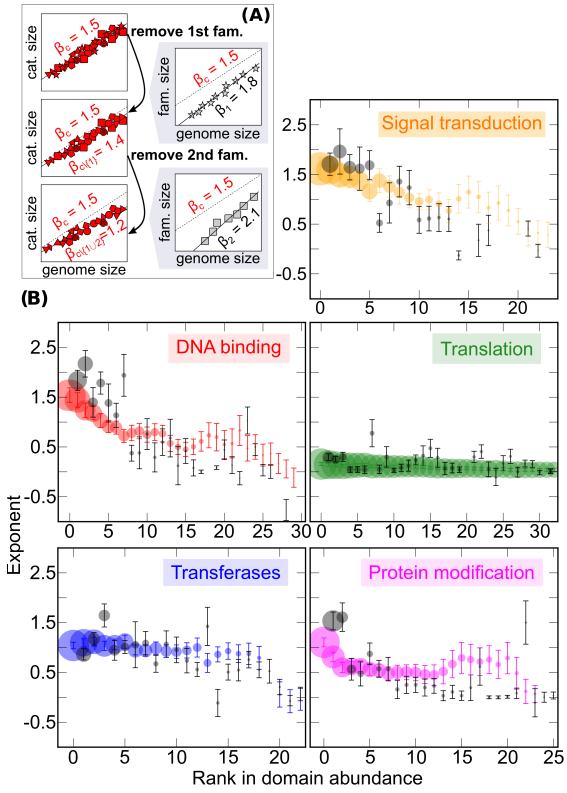

Figure 2.6: Systematic removal of families (ranked by abundance) inside functional categories reveals how individual families build up functional category scaling. (A) Illustration of the procedure. Families belonging to a given functional category are ranked by overall abundance on all genomes and removed one by one from the most abundant. The scaling of the removed family and the remainder of the category is evaluated after each removal. The plots are a stylized example of the first two steps (using values for the category DNA binding). β_c is the category exponent, β_i are family exponents and $\beta_{c\setminus\{i\}}$ are the stripped-category exponents, computed after the removals. (B) Results of this analysis for four functional categories. Grey circles represent the exponents β_i (and their errors) for the scaling law of each family belonging to the functional category (in order of rank in total abundance). Colored circles are the scaling exponents of functional categories without the domains of the i least abundant families. The size of each symbol is proportional to the fraction of domains in the family or family-stripped category. Error bars are uncertainties of the fits.

2.4.1 Super-linear scaling of transcription factors is determined by the behavior of a few specific highly populated families.

We considered, in particular, the case of DNA-binding / transcription factors [11], which are known to exhibit peculiar scaling in bacteria [71, 55]. The abundance of domains in this functional category increases superlinearly (almost quadratically) with the total domain counts [82, 64, 31]. As shown in the first row of Fig. 2.3B, not all the families in this functional category display a superlinear scaling [11], and the collective scaling of the low-abundance families with genome size is much slower (see Fig. 2.6). Fig. 2.6B shows that only the most 5-6 abundant families display a super-linear scaling ($\beta_i > 1$). These are Winged helix DNA-binding domains (34.8% of abundance), Homeodomain-like (23.3 %) lambda repressor-like DNA-binding domains (9.5%) bipartite Response regulators (7.7%) Periplasmic binding protein-like (6.2%), and FadR-like (2.4%). The remaining 16.1% of the DNA-binding regulatory domains follows a clear *sublinear* scaling with genome size (exponent 0.7, see Fig. 2.7).

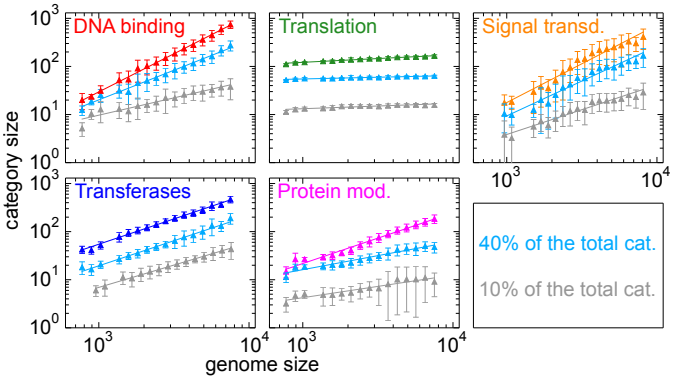

Figure 2.7: Scaling of the abundance of domains belonging to the least abundant families. We progressively removed families from the category in decreasing order of family abundance and represented the scaling of the abundance of the remaining stripped category (y-axis) with genome size (x-axis). The scaling of the stripped category with $\approx$ 40% initial abundance is shown in cyan, while the category with $\approx$ 10% of the initial abundance in grey. The scaling of the original category is shown with a different color (category-specific, as in Fig. 1 and 2 of the Main Text) in each panel.

2.5 Grouping families with similar scaling exponents shows known associations with biological function and reveals new ones.

The above analyses show that the range of scaling exponents of families within the same functional categories is generally wide and that the scaling behavior of some families sensibly deviates from their category. At the same time, functional categories show clear characteristic scaling laws, with well-defined exponents β_c [57]. We, therefore, asked to what extent a range of family scaling exponents β_i is peculiar to a functional category and how this compares to the category exponent β_c. To this end, we grouped families based on their scaling exponents. We then used those groups to test how much specific range of exponents define specific functions by an enrichment test of functional annotations. Let's see in more detail how we performed the analysis.
All families passing the filters described in section 2.2.1 were divided into three groups based on the values of their exponent β_i:

(i) sub-linearly scaling families, $\beta_i \leq 0.6$

(ii) linearly scaling families, $0.6 < \beta_i < 1.4$

(iii) super-linearly scaling families, $\beta_c \geq 1.4$

We used hypergeometric tests to asses over- or under-representation of functional categories in these family groups. Given that F_c is the number of families that belong to the category c, F_{bin} is the number of families in either of the three groups defined above and F_{tot} is the total number of families involved in this analysis, the mean and the variance of the hypergeometric distribution are:

$$\mu_{bin,c} = F_{bin} \cdot \frac{F_c}{F_{tot}} ,$$

$$\sigma^2_{bin,c} = F_{bin} \cdot \frac{F_c}{F_{tot}} \cdot \left(1 - \frac{F_c}{F_{tot}}\right) \cdot \left(1 - \frac{F_{bin} - 1}{F_{tot} - 1}\right) ,$$

For each combination of functional category and family group, the quantity $x_{bin,c}$ is the number of families that lie in the intersection of category c with family group *bin*. The functional category c is under-represented in the group *bin* if $Z_{bin,c} < -1.96$, over-represented if $Z_{bin,c} > -1.96$, where $Z_{bin,c}$ is the Z-score:

$$Z_{bin,c} = \frac{x_{bin,c} - \mu_{bin,c}}{\sigma_{bin,c}} ,$$

In order to prove that the results are independent from the chosen interval of the exponents, we substituted the three groups with sliding intervals of amplitude 0.4 and step 0.1 and repeated the same process. Only intervals with more than 10 families are considered.

The resulting intersection values $x_{bin,c}$ and the significant Z-scores are reported in Table 2.3. This table shows that in most cases functional categories are over-represented in the exponent range where their scaling exponents β_c is found. This confirms and puts in a wider perspective the previously reported strong association between abundance scaling with size and functional annotation. As can be expected from previous results, the functional category Protein Modification is an exception: this category is under-represented in the linear region even though its category exponent is ~ 1.06, since it contains two strongly superlinear families and a bulk of families with sublinear scaling. This strong heterogeneity in scaling exponents is also visible in Fig. 2.6B.

The exponent corresponding to the maximum Z-score defines a representative exponent for each functional category, and can be compared to the exponent β_c measured directly from the plot of category abundance vs genome size (see Fig. 2.8). Interestingly, this analysis also shows that in many cases a single functional category is enriched for multiple groups of families with well-defined exponents, as in the case of the Protein Modification category. The cases of Ion Metabolism and Transport (already discussed), Coenzyme Metabolism and Transport, Redox also shows clear indications of enrichment for two or more exponent groups. For the category Coenzyme Metabolism and Transport this is due to the presence of a single abundant family with scaling exponent close to 2, the acyl-CoA dehydrogenase NM domain-like, whose functional annotation is still not well defined. In the case of Redox, the most abundant families (Thioredoxin-like, 4Fe-4S ferredoxins, Metallo-hydrolase/Oxydoreductase) scale linearly, but there is a wide range of families with exponents between 0.5 and 1, and once again two fairly abundant outlier families with superlinear scaling (Glyoxalase/Bleomycin resistance protein/Dioxygenase, and ALDH-like), both with a fairly wide range of functional annotations.

2.6 The main results of our analysis hold also for PFAM clans

We chose PFAM clans as an alternative database to test the robustness of our results. PFAM clans were annotated on the same scheme of 50 functional categories used for superfamilies, using the mapping of clans into superfamilies available from the PFAM website `http://pfam.xfam.org/clan/browse#numbers` [20]. The scaling laws for functional categories are recovered also for clans (Table A.3) and are consistent with previous results [82, 57, 58, 13, 30, 11]. The following main results were recovered for Pfam clans.

Detailed function	$\beta_i \leq 0.6$	$0.6 < \beta_i < 1.4$	$\beta_i \geq 1.4$	$\beta_c \pm \sigma_{\beta_c}$
Translation	20(4.3)	1(−3.7)	0	0.16 ± 0.03
DNA replication/repair	11	7	0	0.51 ± 0.07
Transport	5	9	1	1.1 ± 0.2
Proteases	7	9	0	0.9 ± 0.1
Protein modification	8	1(−2.3)	2	1.06 ± 0.09
Ion m/tr	11	3	3(−2.2)	1.3 ± 0.1
Other enzymes	29	32	2	1.04 ± 0.06
Coenzyme m/tr	17(2.2)	6	1	0.85 ± 0.09
Redox	4(−3.3)	18(3.1)	2	1.2 ± 0.1
Energy	11	7	0	0.86 ± 0.09
Nucleotide m/tr	16(3.1)	3(−2.5)	0	0.53 ± 0.08
Carbohydrate m/tr	4	8	0	1.0 ± 0.2
Transferases	5	11	1	1.05 ± 0.07
Amino acids m/tr	7	6	0	0.8 ± 0.2
DNA-binding	5	4	4(3.3)	1.5 ± 0.1
Signal transduction	1(−2.7)	5	5(5.0)	1.6 ± 0.2
Unknown function	9	7	0	0.98 ± 0.09

Table 2.3: Family scaling exponents can be associated to specific biological functions. Each cell in the table indicates the number of families that functional categories (rows) share with groups of families whose scaling exponents fall in pre-defined intervals (columns). The table also shows the Z-scores for a standard hypergeometric test (shown in green for over-representation and in red for under-representation, only $|Z| > 1.96$ are shown).

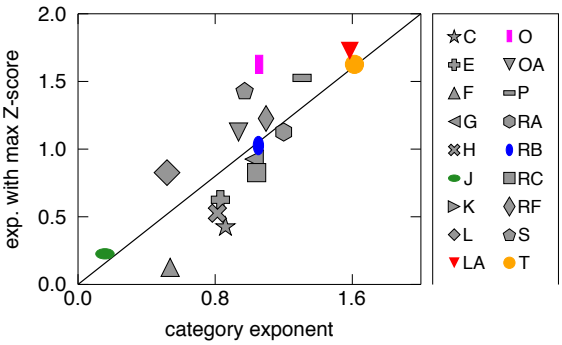

Figure 2.8: Comparison of the category exponent with the exponent corresponding to the maximum Z-score in the enrichment test (see Sec. 2.5). The black line is the $y = x$ line. Correspondence with this line indicates clear association between the functional category and the scaling exponent range. The panel on the right shows the association between symbols and category codes (see Table 2.1 for the corresponding category name).

(i) For each clan, the abundance across genomes scales as a power law of the genome size. Equally to SCOP superfamilies, Pfam clans have individual scaling exponents that may or may not follow the one of the associated functional category (Table. A.3). The fitting method and threshold values are the same used for superfamilies (sec. 2.2.1). 178 clans out of 446 passed the filters and were employed for further analysis.

(*ii*) The heterogeneity (average of the distance between the category exponent and the clan exponent), positively correlates with the category exponent (Fig. 2.9). Functional categories with superlinear scaling tend to be more heterogeneous and, as found for superfamilies, the functional category Signal Transduction is less heterogeneous than DNA-binding, although having the largest exponent. Unlike the case of superfamilies, Protein Modification does not have high heterogeneity score, but the difference in scaling between the (strongly superlinear) outlier family Gro-ES and the remaining ones is observed. For clans, the scaling of Protein Modification is once again strongly biased by the clan "GroES-like superfamily" (20% of the total domains).

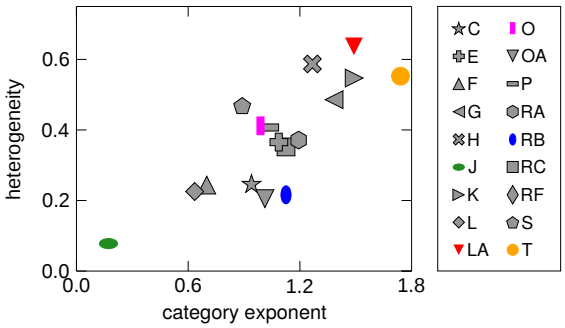

Figure 2.9: Functional categories of Pfam clans with faster scaling exponents contain clans with more heterogeneous scaling laws. Same as Figure 2.5A, for Pfam clans. Heterogeneity is quantified by the mean deviation between the clan scaling exponents and the category exponent. The plot reports heterogeneity scores for different functional categories, plotted as a function of the category exponents. Each symbol corresponds to a different functional category. Only categories with more than 5 clans are shown. The right panel shows the association between symbols and category codes (see Table 2.1 for the corresponding category name).

(*iii*) Either few or most of the clans determine the scaling exponent of the functional category they belong to. Figure 2.10 is coherent with what observed for superfamilies, in particular the functional category of DNA-binding is dominated by one clan (the "Helix-turn-helix" clan) that accounts for 83% of the total domains. As for superfamilies, Signal Trasduction is robust to the progressive removal of families confirming that the presence of dominant clans is not related to the superlinear scaling of the category.

(*iv*) Grouping clans with similar scaling exponents recovers known associations between the category exponent and the biological function (Fig. 2.11).

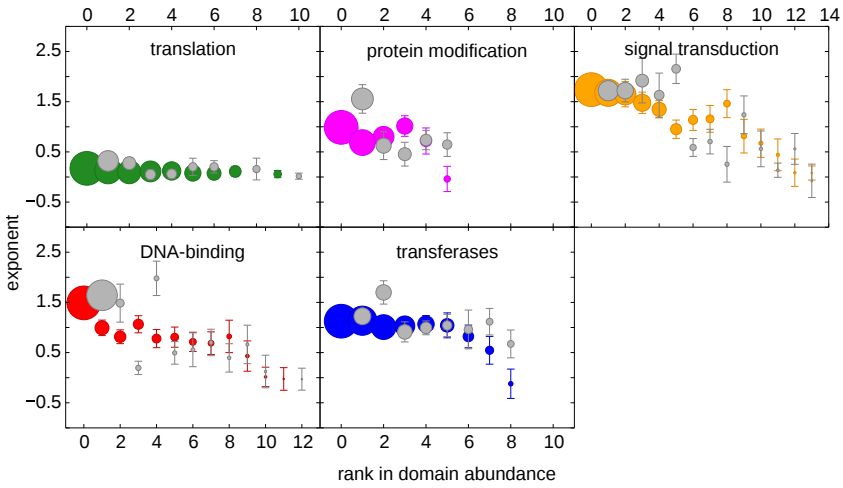

Figure 2.10: Systematic removal of families (ranked by abundance) from functional categories reveals how individual families build up functional category scaling. Same as Figure 2.6, for Pfam clans. Grey circles represent the exponents β_i (and their errors) for the scaling law of each clan (instead of SUPFAM families) belonging to the functional category (in order of rank in total abundance). Colored circles are the scaling exponents of functional categories without the domains of the i most abundant clans. The size of each symbol is proportional to the fraction of domains in the clan or clan-stripped category. Error bars are uncertainties of the fits (see Methods).

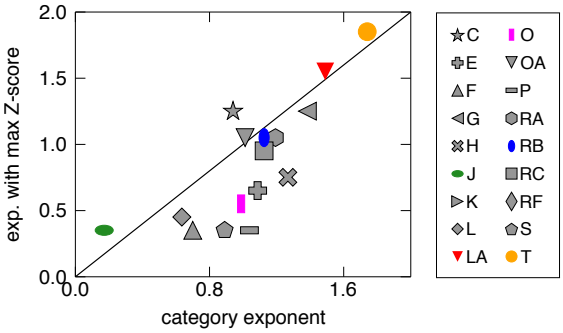

Figure 2.11: Functional enrichment of sets of Pfam clans with similar scaling exponents. Same as Fig. 2.8 for Pfam clans. Comparison of the category exponent with the exponent corresponding to the maximum Z-score in the enrichment test (see Sec. 2.5). Clans are grouped into sliding bins according to the value of their scaling exponent and tested for enrichment against each functional category. The exponent corresponding to the maximal value of Z-score (y-axis) is compared to the category scaling exponent (x-axis). The black line is the $y = x$ line. Correspondence with this line indicates clear association between the functional category and the scaling exponent range. The right panel shows the association between symbols and category codes (see Table 2.1 for the corresponding category name).

2.7 Discussion

Our results gather a critical mass of evidence in the direction of family-specific expansion rules for the families of protein domains found in a genome. Although previous work had focused on individual transcription factor families [11], finding in some cases some definite scaling, no attempts were made to address this question systematically. The scaling laws for domain families appear to be very robust, despite of the limited sampling of families compared to functional annotations (which are super-aggregates of families and hence have by definition higher abundance). In particular, the results are consistent between the different classifications of families we tested (SUPERFAMILY and PFAM, see section 2.6).

Overall, our results indicate that scaling laws are measurable at the family level, and, given the heterogeneous scaling of families with the same functional annotations, families are likely a more reliable description level for the scaling laws than functional annotations. The interpretation of these scaling laws is related to the evolutionary dynamics of family expansion by horizontal transfer or gene duplication, and gene loss [42, 82, 28]. Scaling exponents are seen as "evolutionary potentials" [57], is based on a model of function-specific (multiplicative) family expansion rates. Assuming this interpretation, then our result that these rates may be different for different domain families having the same functional annotation may seem puzzling. Clearly, selective pressure can only act at the functional level, and if two folds were functionally identical, there should be reasonably no advantage selecting one with respect to the other, and doing so at different specific rates. For example, a transcription factor using one fold to bind DNA rather than another one should be indistinguishable from one using a different fold, provided binding specificity and regulatory action are the same.

In view of these considerations, we believe that our findings support a more complex scenario for the interplay between domain families and their functions. Specifically, we put forward two complementary rationalizations. The first is that functional annotations group together different domains whose abundance is linked in different ways to genome size because of their different biochemical and biological functional roles. Such differences may range from slight biochemical specificities of different folds to plain misannotations. This is possible, e.g., with enzymes, where the biochemical range of two different folds is generally different. This observation might be related to the positive correlation we found between the number of EC numbers corresponding to a metabolic domain and its scaling exponent. However, such interpretation might be less likely applicable to, e.g., transcription factor DNA-binding domains, where functional annotation is fairly straightforward [53], but different scaling behaviors with genome size are nevertheless found.

The second potential explanation assumes the point of view where scaling laws are the result of functional interdependency between different domain families [55, 29], then correlated fluctuations around the mean of family pairs should carry memory of such dependency structures [65]. More in detail, there may be specific dependencies connecting the relative proportions of domains with both different and equal functional annotations that are present in the same genome, which might determine the family-specific behavior [30]. While further analysis is required to elucidate these trends, we believe that gaining knowledge on functional dependencies would be an important step

to understand the functional design principles of genomes. It is not possible at this stage to distinguish between these two explanations, and we surmise that they may both be relevant to explain the data.

Of notable importance is the case of the superlinear scaling of transcription factors, which has created notable debate in the past [72, 55]. For the first time, we look into how this trend is subdivided between the different DNA binding domains [53]. Our analysis indicates that the superlinear scaling is driven by the few most abundant superfamilies (mostly winged-helix, homeodomain, lambda repressor). However, the remaining 10-20% of the functional category gives a clear sublinear scaling with genome size, which emerges beyond any sampling problems. We speculate that these other regulatory DNA-binding domains may be functionally different or behave differently over evolutionary time scales. Hence, the scaling of transcription factors with size in bacteria is driven by a small set of domain families with scaling exponent close to two, which take up most of the abundance, but does not appear to be peculiar of *all* transcription factors. A "toolbox" model considering the role of transcription factors as regulator of metabolic pathways and the finite universe of metabolic reactions [55, 64] predicts scaling exponents close to two for transcription factor families. According to our results, such model should be applicable to the leading TF families. Interestingly, the heterogeneity in the behavior in transcription factor DNA-binding domains is much higher than that of the other notable superlinear functional category, signal transduction, where removal of the leading families does not significantly affect the observed scaling of abundance with genome size. Given the clarity and uniformity of the scaling exponent, we speculate that possibly a toolbox-like model may be applicable to understand the overall scaling of this category.

Other categories clearly contain multiple sets of families with coherent exponents or single outlier families. In some cases, two main groups of families with different scaling behavior clearly emerge, and higher observed scaling exponents may be related to a wider range of functional annotations. We propose that such easily detectable trends can be used to revise and refine functional annotations of protein domains. Such functional annotations are currently largely curated by humans, and based on subjective and/or biased criteria. The analysis of family scaling gives an additional objective test to define the coherence of the families that are annotated under the same function. While yet-to-be-developed automated inference methods based on our observations could serve this purpose, the quantitative scores defined here already provide useful information. The heterogeneity of a functional category is an indication of how likely that group of domain families follows a coherent expansion rate over evolution. The enrichment scores for sets of families with a given range of scaling exponent helps to pinpoint the sets of families within the functional category that expand coherently with genome size.

Chapter 3

Dependency networks shape frequencies and abundances in component systems

3.1 Introduction

3.1.1 The emergence of universal regularities in empirical component systems may be the effect of underlying dependency structures of the components.

The partitioning of a system into components is a general architectural pattern present in many physical, biological, and artificial systems. For example genomes can be regarded as sets of genes, operating systems such as Linux can be thought of as sets of packages, texts can be analyzed as sets of words. Modular representations of component systems emerge in diverse fields (e.g. quantitative geography [8, 51], linguistics [22], software [65]. Such a "toolbox" structure [8, 22, 23, 51, 55, 65] extends the classic partitions that are a core subject of statistical mechanics, such as equilibrium statistics of particles in energy states [47] or non-equilibrium occupancy, for example in driven diffusive systems [18] or general duplication-innovation models [3, 16].

A variety of quantitative laws have been uncovered studying different systems, some of which are system-specific, while some others are common to multiple systems. In some cases, scale-invariance (and universality) emerges as a consequence of criticality [77], either due to evolutionary tuning or self-organization, as is well understood via the renormalisation group in statistical mechanics. However, more and more often scale-free features are recognized as non-universal consequences of collective behavior, non-linear dynamics, preferential attachment, etc. [76]. Prominent examples are the followings:

- *Zipf's law* that concerns the distribution of component frequencies [67];

- the power-law distribution of component occurrences;

- the sublinear scaling of the number of component classes with system size, often referred to as Heaps' law [34].

In Chapter 2 we saw that the concept of component systems has helped unveiling emergent "laws" or quantitative invariants pointing to relevant underlying evolutionary and architectural properties of genomes [30, 49, 55, 65, 82], as well as to the value of an "emergent" description of genome architecture [42]. Due to the interest in field-specific studies, it is important to gain a theoretical understanding of the traits emerging from the common architecture of component systems, and of their implications in a Interesting results have been obtained in the context of history-dependent processes, where the state space changes (co-evolves) with the realizations [32]. Two divergent mechanisms have been proposed, where the state space expands [79] or reduces [15, 14] while it is explored by the system.

Dependency structures (DS) have emerged recently as a promising framework for the rationalization and organization of the regularities observed in systems lying outside the traditional scope of statistical mechanics [65]. Dependency structures have been proposed in various contexts and forms, and have helped achieving remarkable results, for instance in the scope of preference prediction [33], or for addressing causality in financial data [40].

A DS is a directed graph (most often, but not necessarily, acyclic), whose nodes are the components (e.g. Linux packages, or genes [65]) and whose links are the dependency relations occurring between them. A component depends on another one if it is not functional unless the latter is present. A *realization* in such a component system is then constructed as the choice of a node and all its direct and indirect dependencies. This simple model links quantitative laws in the modular representations to topological properties of the DS (and hence to the evolutionary processes sculpting it). For instance, a broad ensemble of DSs has the property that the number of total dependencies of each node is scale free in the thermodynamic limit (notice that this is a weaker condition than the power-law distribution of degrees, i.e., of direct dependencies). This topological property explains the fat-tailed distribution of component occurrences across realizations, both in genomes and operating systems [65].

This model constrains the components to have binary abundances in its realization, i.e., to either be present in one copy or to be absent. Such a description is expected to be accurate for some components (e.g., for software packages) but is a rough approximation for those systems where components appear with non-negligible abundances (e.g., coarse-grained evolutionary families of genes such as superfamilies, and words in a text).

This chapter extends the model proposed in [65] to a case where components appear with non-trivial abundances. This allows us to explore how dependency structures affect abundance-related features, such as Heaps' and Zipf's laws.

3.2 Model: description of the dependency structure and the algorithm that defines a realization.

This section discusses the definition of the model. Let $\mathcal{U}$ be the set of all unique components (the *universe*), and let $U = |\mathcal{U}|$ denote its cardinality. A *realization* of this component system is a set $r = \{c_i\}$ of components $c_i \in \mathcal{U}$, with $i = 0, \dots, N_r$, where N_r is the *size* of the realization. Such a simplification separates the relational constraints existing between the components, such as dependency and incompatibility, from their functional correlations, such as synergy, co-occurrence, interchangeability, conflict, and so on. The constraints are realized by a network of dependencies, as we explain below. For what concerns the functional correlations, our model is the simplest null model, where no correlations between components are dictated, other than those arising from the dependencies.

A *dependency structure* is a directed acyclic graph $\mathcal{G}$ on $\mathcal{U}$, which encodes the dependencies between the components. An edge $i \to j$ between two nodes i and j represents the relation "i depends on j". A component i is said to depend on another component j if i is not functional without j. Such a relation can be more or less strict depending on the system; for instance it is enforced in software operating systems, where a package cannot function unless all its dependencies are installed, but not in metabolic networks, where alternative pathways can be followed to the same metabolite [23, 55]. We assume here strict unbroken dependencies. Notice that acyclicity of $\mathcal{G}$ is not stringently necessary; however, as will be clear in the following, a cycle in $\mathcal{G}$ would behave as a single node in the model.

The topology of the dependency structure is conceptually separated from the procedure that generates realizations satisfying the dependency constraints. Here we use the DS introduced in [65] and define a novel method to build the realizations. Specifically, as sketched in Fig. 3.1A, the growth process that creates the dependency network is a very simple incremental node-addition process that generates structures with power-law distributed sizes of a nodes' direct and indirect dependencies (such property is crucial to reproduce the Zipf's law, see section 3.3). Let us fix an average outdegree $D \geq 1$, i.e., an average number of direct dependencies that a given component has. Starting with an initial graph consisting of a single node, the full graph is built node by node, by attaching the new node to $d + 1$ randomly chosen existing nodes (possibly with repetitions), where d is a Poissonian random variable of mean $D - 1$. The process is stopped when the network reaches the predetermined size U. A graph grown with these rules is directed and acyclic, and hence a good dependency structure, as can be seen by labelling each node by the time $t = 1, \dots, U$ it was added to the network, and noticing that there can be no links $t \to t'$ with $t < t'$.

Once a dependency structure $\mathcal{G}$ is established, realizations of the model, i.e., sets of components, are generated by the following procedure (see also Fig. 3.1A). Before explaining the details of the algorithm, it is crucial to state the following definitions:

- Given a node c, $\wedge(c) \subseteq \mathcal{U}$ is defined as the set of all nodes c' such that there exists at least one directed path in $\mathcal{G}$ starting from c and arriving at c'. We will call the set $\wedge(c)$ the *forward cone* of c.

- We define the *backward cone* $\vee(c)$ of c as the set of all nodes c' such that there exists a path from c' to c.

In other words, $\wedge(c)$ is the set of all components required be the (direct and indirect) dependencies of c, whereas $\vee(c)$ is the set of the nodes that depend (directly or indirectly) on c.

Let us fix a positive integer k, which represents the number of "precursor" components determining a realization. The k precursors $\{p_j\}$, $j = 1, \ldots, k$, are chosen randomly and independently among the nodes of $\mathcal{G}$. Then the corresponding realization is produced by taking all components belonging to the forward cones of the precursors. To complete the model specification one needs a rule to choose the multiplicities of the components. We allow a component belonging to multiple cones to appear in multiple copies. Let us imagine that precursors are added one at a time to the realization. At the j-th step the existing realization r_{j-1} (possibly empty, when $j = 1$) is extended, by the addition of elements from the cone $\wedge(p_j)$ of the precursor p_j:

$$r_j = r_{j-1} \cup \Delta_j, \quad \Delta_j \subseteq \wedge(t_j). \tag{3.1}$$

The choice of the incrementing set Δ_j must be done so as to satisfy the dependency relations, i.e., $r_j \supseteq \wedge(t_j)$. Doing this at every step ensures that the final realization r_k will not have any unsatisfied dependency. Other than that, Δ_j is in principle unconstrained, and it may be a random variable even at fixed p_j and r_{j-1}. Here we make the simplest choice $\Delta_j = \wedge(t_j)$. This makes the process Markovian, in the sense that $r_j \setminus r_{j-1} \equiv \Delta_j$ is independent of r_{j-1}. With a slight abuse of notation, we will write $\wedge(t)$ and $\vee(t)$ for the forward and backward cones of the t-th node. The case $k = 1$, when a realization is specified by a single precursor, reproduces the model of [65].

An advantage of this model is its analytical tractability. The forms of Zipf's law and the distribution of component occurrence are the same as those in the binary model, and the mean-field analysis is the same as that in [65]. The main additional output of our extension is a non-trivial Heaps' law, which is derived analytically in section 3.4.

3.3 Our positive model recovers the empirical regularities of component systems, namely the Zipf's law and the Heaps' law.

Given a set of N realizations of a component system, the "popularity" of a component i can be measured in two ways: by its *abundance* a_i and by its *occurrence* o_i. The relative abundance counts the number of times that i appears in all realizations (with multiplicities):

$$a_i = \frac{1}{kN} \sum_r \sum_{c \in r} \delta_{c,i}. \tag{3.2}$$

In the model, the maximum abundance of a component i corresponds to drawing i each time a cone is selected, for each realization. In such a case, the double sum in (3.2) is kN. Therefore, the abundance a_i is normalized so that $0 \leq a_i \leq 1$. It is important to stress that the *relative* abundance

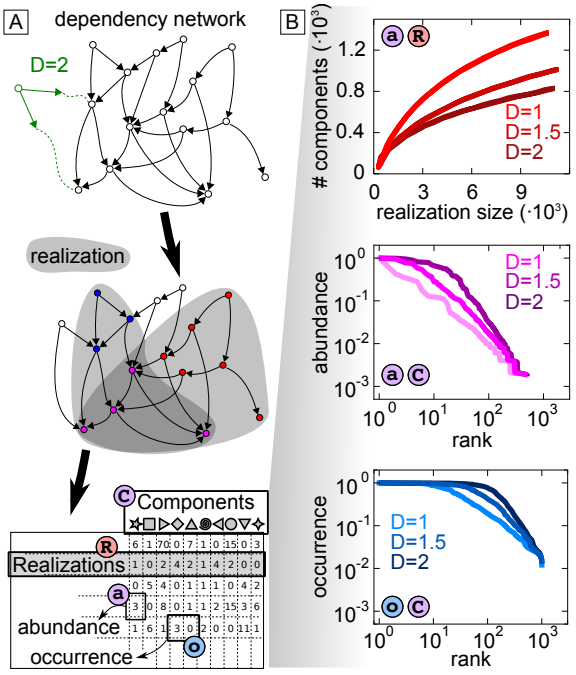

Figure 3.1: Illustration of the model and main observables. (A) The model is composed of three steps. The first step creates a dependency structure by an incremental node-addition process with mean out-degree D. The second step builds realizations drawing at random k precursors and taking all components belonging to their forward cones. The final step assigns multiplicities to the components. The model establishes that the selected components have abundance equal to the number of forward cones that contain them. (bottom panel) Each realization constitutes the row of a matrix in which columns are components, therefore the matrix element m_{ij} is the multiplicity of component j in realization i. The component's abundance is the multiplicity, whereas the occurrence refers to the presence or absence of the component. The system constituted by an ensemble of realizations features known empirical laws: sublinear scaling of the number of distinct components with realization size (top panel), zipfian distribution of components abundance (central panel) and power law distribution of components occurrence.

is an intensive quantity, not to be mistaken for the *absolute* abundance that coincides with the components' multiplicity (the latter is the definition of abundance used in Chapter 2).

The mean occurrence o_i measures the fraction of realizations containing the component, regardless of its abundance:

$$o_i = \frac{1}{N} \sum_r \left[1 - \prod_{c \in r} (1 - \delta_{c,i}) \right]. \tag{3.3}$$

With this definition, the mean occurrence is normalized so that $0 \leq o_i \leq 1$.

3.3.1 The analytical derivation of the components abundance distribution matches simulations and satisfies the Zipf's law

Zipf's law is an empirical law about the rank-frequency relation of components. It states that the frequency of any components across realizations is inversely proportional to its rank. This behavior was first identified in linguistics, but appears to be a feature of very diverse systems, all of which have a component-realization structure. Many attempts have been made to explain the emergence of such regularity [25], in the spirit of the paper by Pang and Maslov [65] we show here that our model generates component with a zipfian abundance distribution as a consequence of the dependency structure.

The Zipf relation, i.e., the rank plot of the abundance, is expected to be independent of the number of cones k (at least for large systems). In fact, the abundance a_i of a given component in

N realizations constructed with k cones each has the same distribution as that in kN single-cone realizations, since the choices of the cones are independent. a_i can be estimated as the probability of choosing a cone that contains i, which is proportional to the size $|\vee(i)|$ of the backward cone of i:

$$a_i = \frac{|\vee(i)|}{U}. \tag{3.4}$$

Let us call rank(i) the rank of component i when all components are ranked by their abundance. Following [65], an approximate relation can be derived between $|\vee(i)|$ and rank(i), which will allow to obtain an analytical estimate of Zipf's law. The t-th node in the network (the one added at the t-th step of the construction, when a network of size $t-1$ has already been generated) has approximately $(U/t)^D$ nodes that depend on it. This result can be obtained by writing an equation based on the observation that the backward cone of the t-th node is the union of the backward cones of all the nodes that, at later times t', will directly attach to the t-th node. Neglecting the intersections between these cones allows to write the recursion

$$|\vee(t)| = 1 + \sum_{t'=t+1}^{N} \frac{D}{t'} \left|\vee(t')\right|, \tag{3.5}$$

where the factor D/t' estimates the probability that the t'-th node attaches to the t-th node. By approximating the sum by an integral and taking a derivative with respect to t, one obtains a differential equation that is solved by $|\vee(t)| = (U/t)^D$.

For small t, however, $(U/t)^D$ is greater than the size of the network U. In fact, the relation can hold only down to a cutoff $t_{\min}$, which can be estimated by the condition that the whole network depends on the $t_{\min}$-th node, i.e., $(U/t_{\min})^D = U$, which gives $t_{\min} = U^{1-1/D}$. For any node below $t_{\min}$, the size of its backward cone is $\approx U$:

$$|\vee(t)| \approx \begin{cases} U & t < U^{1-1/D} \\ (U/t)^D & t \geq U^{1-1/D} \end{cases} \tag{3.6}$$

Equations (3.4) and (3.6) imply that if node i is the t-th node in the network growth process, then $t = \text{rank}(i)$. (This identification does not hold for the first $U^{1-1/D}$ components, but this does not influence the result since the size of their backward cones are equal in this approximation.) Therefore, one obtains

$$a_i \approx \begin{cases} 1 & \text{rank}(i) < U^{1-1/D} \\ \text{rank}(i)^{-D} U^{D-1} & \text{rank}(i) \geq U^{1-1/D}. \end{cases} \tag{3.7}$$

This relation has the form of a Zipf power-law (with exponent $-D$) with an initial "core" consisting of $U^{1-1/D}$ components having similar abundances.

Figure 3.2A compares the analytical form (3.7) with the results of simulations, showing good accord, especially in the behavior of the fat tail. The transition between the core and the tail, instead, is less sharp than predicted. This is tied to the fact that the relation $|\vee| = (U/t)^D$ starts to break down before reaching U, and saturates more smoothly than in the approximation made above. Importantly, the relation between rank and relative abundance does not depend on the number of cones k, in agreement with the above prediction.

3.3.2 The power-law distribution of components occurrence is a "null" result of our model

The occurrence-abundance relation predicted by our model turns out to be universal (or "null"), meaning that it is insensitive to the explicit form of Zipf's law, to the detailed structure of the network, and even to the size U of the component universe. In fact, we show here that a simple probabilistic argument gives a relation that is consistent with simulations of the full model.

In the limit of large N, we can assume that the occurrence of a component i is equal to the probability of choosing i at least once in a single realization: $o_i = 1 - (1 - a_i)^k$, hence

$$a_i = 1 - (1 - o_i)^{1/k} . \tag{3.8}$$

For realizations with a single precursor ($k = 1$) abundance and occurrence are equal. While k increases, more and more components (with larger and larger occurrence) assume small abundances. In the large-k limit, all components have zero (relative) abundance, except those with occurrence 1. Figure 3.2C shows that a scatterplot of abundance versus occurrence in simulations perfectly matches the theoretical curve (3.8). The figure shows results for a single choice of D and U, but we verified that these parameters have no effect on the curves.

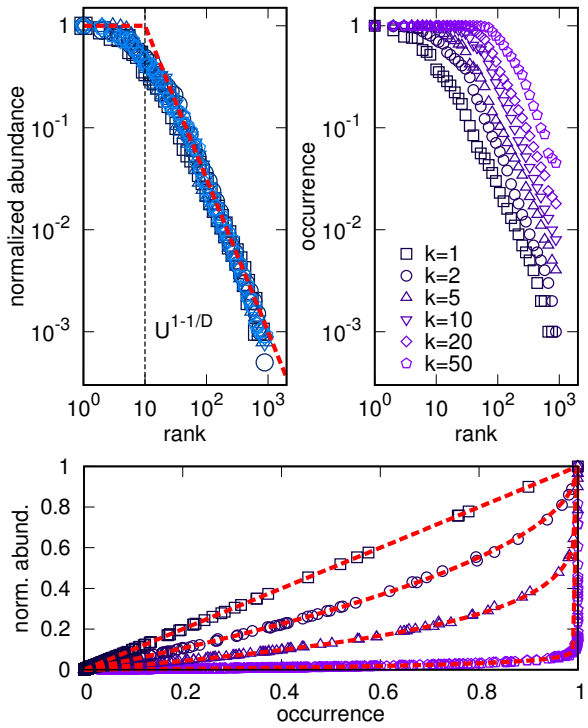

Figure 3.2: Simulations match the analytical form of the Zipf's law and of the abundance-occurrence relation. From the top-left corner: the rank-plot of components abundance, the rank-plot of components occurrence and the scatterplot of components abundance (y-axis) versus components occurrence (x-axis). Colored symbols represents simulations with varying number of precursors k, the dashed line is the analytical prediction. The simulation parameters are $U = 1000$, $D = 1.5$ and the number of realizations is fixed to 1000.

3.4 The analytical mean-field expression of the Heaps' law matches the results of numerical simulations of the model.

We now set out to estimate analytically the Heaps' law from the model. The calculation of the number $F(N)$ of unique components in a realization of size N can be performed in a mean-field approximation, where the correlations between nodes are neglected. We consider a process where a realization is generated by extracting N nodes independently. The probability

$$p(t) = \frac{1}{\Omega} |\vee(t)| \tag{3.9}$$

of drawing the node t is proportional to the size $|\vee(t)|$ of the node's backward cone. In a continuous approximation, the normalization Ω can be fixed by the condition $\int_0^\infty p(t)\, \mathrm{d}t = 1$, which yields

$$\Omega = U \frac{DU^{1-1/D} - 1}{D - 1}. \tag{3.10}$$

Note that $\Omega > U$ whenever $U > 1$ and $D > 1$. Let $p_1(t, n)$ be the probability that the t-th node in the network is drawn for the first time when the system being constructed has size n:

$$p_1(t, n) = p(t) \left[1 - p(t)\right]^{n-1}. \tag{3.11}$$

A mean-field estimate of F can then be obtained as

$$F(N) = \sum_{n=1}^{N} \sum_{t=1}^{U} p_1(t, n) \approx \int_0^U \mathrm{d}t \sum_{n=1}^{N} p_1(t, n) \tag{3.12}$$

The geometric sum in n gives simply the probability $1 - [1 - p(t)]^N$ that the t-th node has been drawn at least once after N steps. The mean-field expression for Heaps' law is then given by the following integral:

$$\begin{aligned} F(N) &= \int_0^U \mathrm{d}t \left\{1 - [1 - p(t)]^N\right\} \\ &= U - U^{1-1/D} \left(1 - \frac{U}{\Omega}\right)^N - \mathcal{I}(N), \end{aligned} \tag{3.13}$$

where the first term (U) comes from the integral of 1, and the second and third terms are the contributions of the two regions in (3.6). The remaining integral

$$\mathcal{I}(N) = \int_{U^{1-1/D}}^{U} \mathrm{d}t \left[1 - \frac{1}{\Omega} \left(\frac{U}{t}\right)^D\right]^N \tag{3.14}$$

can be evaluated with the change of variables $z = (U/t)^D/\Omega$, which gives

$$\mathcal{I}(N) = \frac{U}{D} \Omega^{-1/D} \int_{1/\Omega}^{U/\Omega} (1 - z)^N z^{-1-1/D} \mathrm{d}z. \tag{3.15}$$

By remembering that the primitive of $(1-z)^{\alpha}z^{\beta}$ is $z^{\beta+1}\ {}_2F_1(-\alpha,\beta+1,\beta+2,z)/(\beta+1)$, where ${}_2F_1$ is the Gauss hypergeometric function, one finally obtains

$$\begin{aligned} F(N) = U &- U^{1-1/D}\left(1-\frac{U}{\Omega}\right)^N \\ &- {}_2F_1\left(-N,-\frac{1}{D},1-\frac{1}{D},\frac{1}{\Omega}\right)U \\ &+ {}_2F_1\left(-N,-\frac{1}{D},1-\frac{1}{D},\frac{U}{\Omega}\right)U^{1-1/D} \end{aligned} \tag{3.16}$$

Fig. 3.3 shows that the analytical mean-field expression (3.16) nicely matches the results of numerical simulations of the model.

3.4.1 The analytical expression of the Heaps' law shows three different regimes

If a realization is constructed by incremental addition of randomly chosen components, one expects $F(N)$ to be approximately linear for small N, as it is unlikely to draw the same component twice. Intuitively, the probability to do so increases with N, up to a point where approximately all components in the universe will have been included, and $F(N)$ will saturate to U. This behavior is clearly visible by plotting $F(N)$ in log-log scale (see Fig. 3.3A). There emerge three distinct regimes: a linear increase for small N, a saturation to U for large N, and an intermediate regime where the sub-linear increase of $F(N)$ appears to be well described by a power law. Two transition points can be identified, N_c and N_s, respectively at the crossover between the linear and the sub-linear regimes, and at the onset of saturation. A few analytical estimates about the different regimes and observations are possible.

It is clear from expression (3.13) that, since $p(t) > 0$ for a finite universe,

$$\lim_{N\to\infty} F(N) = U. \tag{3.17}$$

This is a consequence of the definition of the model, whereby $F(N)$ is monotonic by construction and $F(N) \leq U$. However, this limit is not apparent from the final formula (3.16). What happens is that the (essential) singularities of the two hypergeometric functions cancel out in the large-N limit. This makes it difficult to compute values of $F(N)$ numerically in this regime (see below).

An estimate of the point N_s where the saturation regime sets in can be obtained from (3.13). The term with $(1-U/\Omega)^N$ is significantly different from zero when $U/\Omega \lesssim 1/N$, i.e., when $N \lesssim \Omega/U$. The integral $\mathcal{I}(N)$, instead, can be evaluated for large N in a saddle point approximation. The integrand [Eq. (3.14)] attains its minimum at $t = U$, where it is equal to $(1-1/\Omega)^N$; hence, it is significantly different from zero when $N \lesssim \Omega$. Therefore, both N-dependent terms in (3.13) are negligible when $N \gtrsim N_s = \Omega$, where Ω is given by (3.10).

The small-N behavior at finite U can be obtained in principle from Eq. (3.13) as well, by expanding in N before performing the integral in $\mathcal{I}(N)$. However, it is easier to analyze the onset

of sublinearity in the large-U regime, by expanding Eq. (3.16) in powers of $1/U^{1-1/D}$. This can be done by using the definition of the hypergeometric function:

$$ {}_2F_1(a,b,c,z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k k!} z^k, \tag{3.18} $$

where $(\alpha)_k = \alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1)$ is the Pochhammer symbol. The two terms of the form ${}_2F_1(...)$ in (3.16) can be expanded in powers of z; Eq. (3.18) shows that the term of order z^j in such an expansion is a polynomial of order j in N. The same property holds for the small-z expansion of the first N-dependent term in (3.16), of the form $(1-z)^N$. It is then easy to see, by keeping track of the analytical and non-analytical powers of U, that in the limit $U \to \infty$ the only non-vanishing term is N, and Heaps' law reduces to the identity

$$ \lim_{U\to\infty} F(N) = N. \tag{3.19} $$

A linear onset is expected for small N even when U is finite. Performing explicitly the expansion to first order in $U^{-1+1/D}$ yields

$$ F(N) \approx N - \frac{1}{2} N(N-1) \frac{2(D-1)^2}{D(2D-1)} U^{-1+1/D}. \tag{3.20} $$

The crossover point N_c separating the linear and sublinear regimes can be estimated by the point where (3.20) reaches its maximum:

$$ N_c = \frac{D(2D-1)}{2(D-1)^2} U^{1-1/D}. \tag{3.21} $$

This expression is expected to become inaccurate when $D \approx 1$ (where in fact it diverges), because all terms of order $U^{-j+j/D}$ with $j > 1$, which are neglected in (3.20), approach constants for $D \to 1$.

In order to bring out the transition points more sharply from the data, one can plot the *effective exponent*

$$ \gamma_{\text{eff}}(N) = \frac{\mathrm{d}\log F(N)}{\mathrm{d}\log N} \tag{3.22} $$

which is easily computed from numerical data as a discrete derivative. $\gamma_{\text{eff}}(N)$ measures the apparent exponent that is obtained by approximating the function $F(N)$ locally by a power-law $N^{\gamma_{\text{eff}}}$. Figure 3.3 shows the effective exponent for a range of values of U. For small U, the regimes are somewhat intertwined, and no sharp transitions appear. For larger U, γ_{eff} shows three plateaux, corresponding to $\gamma_{\text{eff}} = 1$, $\gamma_{\text{eff}} = 0$, and an intermediate value $\gamma_{\text{eff}} = \gamma$.

The figure also shows that the transition points computed above, i.e., N_c given by (3.21), and

$$ N_s = U \frac{DU^{1-1/D} - 1}{D-1}, \tag{3.23} $$

are reasonable estimates of the sizes where the two regime shifts occur. Unexpectedly, the estimate of N_s turns out to correspond to an approximately U-independent value of the effective exponent.

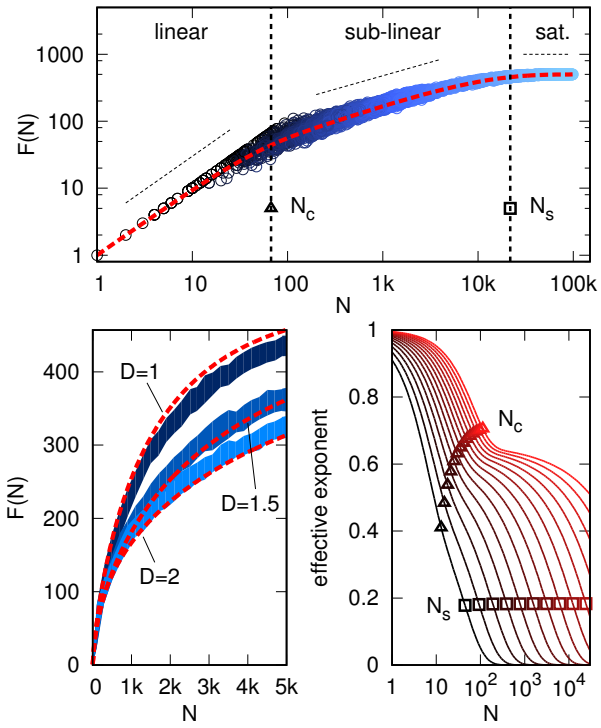

Figure 3.3: Simulations of the model reproduce the characteristic three-regimes structure of Heaps' law. (top) This plot highlights the characteristics three regimes of Heaps' law generated by the model. The colored circles result from the simulations, lighter shades of blue correspond to higher values of precursors k. The red dashed line is the analytical prediction. The scale is logarithmical for both axis. (bottom left) The plot shows the function $F(N)$ in lin-lin scale for three values of the mean out-degree D: 1 (dark blue), 1.5 (blue) and 2 (light blue). The area of the solid curves represents the 90% variability interval. The red dashed lines are the analytical predictions. (bottom right) The effective exponent is plotted against the realization size. In this plot the mean out-degree is fixed at $D = 2$. Parameters: $U = 500$, k ranges from 1 to 3000 and the number of realizations is 13000.

3.4.2 The stretched-exponential saturation is a remarkably good approximation of the simulated data.

As pointed out above, the asymptotically flat behavior of $F(N)$ results from the cancellation of two infinities in the analytical formula. This subtlety makes it numerically challenging to evaluate $F(N)$ especially for large U and N. Such a difficulty prevents the use of Eq. (3.16) for fitting against empirical data. However, the analytical expression (3.15) suggests a simple phenomenological expression, which can be useful for fitting. Since the integration variable z is small for large U, one can attempt to approximate the integrand in $\mathcal{I}(N)$ by $z^{-1-1/D}\exp(-zN)\mathrm{d}z$. In this form, the integral is similar to a representation of the stretched exponential function $\psi_{\gamma,a}(x) = \exp(-ax^{\gamma})$ in terms of exponential decays

$$\psi_{\gamma,a}(x) = \int_0^{\infty} P_{\gamma,a}(z)e^{-zx}\mathrm{d}z \tag{3.24}$$

The asymptotic behavior of $P_{\gamma,a}(z)$ is known to be

$$P_{\gamma,a}(z) \sim z^{\gamma+1} \tag{3.25}$$

for large z, and an exponential decrease for small z [37]. This suggests the following phenomenological expression:

$$F_{\mathrm{ph}}(N) = U\left[1 - \exp\left(-aN^{\gamma}\right)\right] \tag{3.26}$$

Figure 3.4 shows that the stretched-exponential saturation, Eq. (3.26), is a remarkably good approximation of the simulated data. The log-linear scale reveals that the accord is tight on the whole range of N. However, the phenomenological expression fails at capturing the transient linear increase at small sizes (see Sec. 3.4.1). Indeed, the small-N behavior of $F_{\rm ph}$ is $F_{\rm ph}(N) \sim UaN^{\gamma}$. Interestingly, if one extracts γ by matching the large-z power-law scaling in Eq. (3.25) with the factor $z^{-1-1/D}$ in Eq. (3.15), one obtains $\gamma = 1/D$. This same exponent can be derived in the framework of the Zipfian ensemble computations by a simple scaling argument by considering a pure power-law behavior of $F(N)$. Note, however, that the integration range in (3.16) is very different from the one in the integral representation (3.24) of $\psi_{\gamma,a}$, that is $(0, \infty)$. As a consequence, the fitted exponents γ can deviate from the simple scaling relation $\gamma = 1/D$. Altogether, it is quite surprising that the stretched exponential can be such a good approximation.

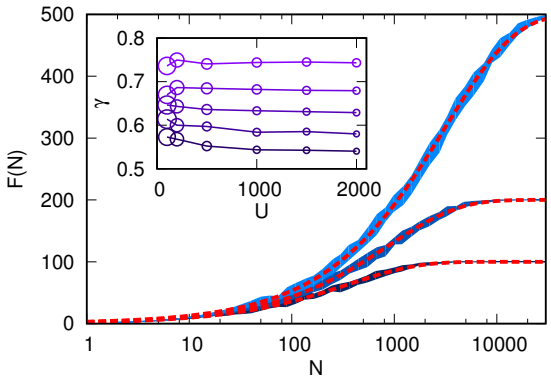

Figure 3.4: The stretched exponential is a good approximation of the Heaps' law generated by the model. (main panel) The plot shows the number of components ($F(N)$ on the y-axis, linear scale) as a function of the realization size (n on the x-axis, logarithmic scale). $F(N)$ has been calculated for three different values of U . Simulations are the solid curves (dark blue $U = 100$, blue $U = 200$ and light blue $U = 500$) where the area of the curve reflect the 90% variability interval. The dashed red lines are the analytical predictions. The mean out-degree is fixed at $D = 1.5$ and the number of precursors k goes from 1 to 5000. (inset) The inset shows the variability of the stretched exponential exponent γ with the size of the universe U. Dots in darker shades of purple represent increasing values of D.

3.5 Conclusion

In conclusion, this Chapter provides the simplest generative mechanism of realizations of non-binary component systems from a dependency structure. The model extends the one proposed in [65] to the case of components with non-binary abundance, and it is the only possible extension that is not history-dependent. Indeed, other possible generalizations need to specify how the dependency cone of the next move intersects with previously generated cones, making each move depend from all previous ones. The additive choice taken here, where each selected component determines the addition of one element to itself as well as to all the components in its dependency cone, provides a minimal model that is memoryless, and therefore still accessible analytically.

The main results are derivations of the universe components abundance (corresponding to Zipf's law [67]) and the (corresponding to Heaps' law [34]). The distribution of components abundances a power-law tail with an initial core, as in the case of the model of ref. [65]. However, in our case the situation is more complex, as the distribution of component abundance and the distribution

of shared components do not coincide, due to the non-binary nature of the occurrence of components in realizations []. We also found that the scaling of component number with the realization size (analogous to the scaling of family number with genome size in genomes) is sublinear, and its analytical form, which we were able to derive, is well approximated with a stretched exponential. Both analytical calculations and simulation show the characteristic three regimes of Heaps' law.

Since a wide variety of systems can be represented as component-systems ensembles, the model defined here has a general applicability, but the interesting case for the scopes of these thesis is that of genomes. For genomes, a dependency structure represents the recipes binding the functional roles of different protein families, thereby determining their usefulness in the same genome. For example, a gene could depend on another one if it is found downstream in the same metabolic pathway [65]. The topology of such dependency has not been fully characterized. Likely, it comprises both feedforward and feedback structures, as well as non-directed exclusion principles (whereby a domain might not be necessary or useful if another one is present). Therefore, it is unlikely that its structure is similar to the simple random graphs considered in this study. Future investigations could aim at defining more stringently from data the minimal features of a dependency structure that could realistically describe genomes. This could be inferred by the correlation structure of domain abundances from sets of entirely sequenced genomes.

The other simplifying hypothesis that needs to be discussed for the case of genomes is the rule used for duplicating domains, which assumes that the whole dependency cone of a chosen component increases its abundance by one. As explained above, this rule makes the process memoryless, and makes analytical calculations possible. However, in the empirical situation, it is possible (and likely) that not all gene families in the dependency cone of a chosen one need to double, and likely the pre-existence of domains in the cone plays a role. For the above two reasons, it is difficult to compare the model results directly to data. However, comparing the scaling of the theoretical prediction of the number of distinct families with genome size, we clearly noticed that a fit with a stretched exponential (or with the direct analytical prediction of the model) works much better than a power law or logarithmic growth. The joint prediction of Heaps' law and the universe distribution of components is more difficult to reproduce, and requirer more precise knowledge of both the evolutionary rules and the dependency structure of the empirical system.

Chapter 4

Signature of gene-family scaling laws in microbial ecosystems

4.1 Introduction

Metagenomics is defined as the direct analysis of genomes contained within an environmental sample [78]. Since the DNA analyzed comes directly from the uncultured microbial community, metagenomics has given us an unprecedented view on the diversity, composition, and dynamics of microbial ecosystems [68, 12, 56, 73, 9]. It has brought considerable insight into intra-species interaction in varying habitats [86], which is not possible in clonal communities, therefore establishing its complementary role to single-organism genome studies. The access to uncultured ecosystems discloses new taxa or protein families [60] that reference databases can not capture.

However, there is still a considerable imbalance between the large amount of available data and the quantitative grasp we are able to consolidate on these systems. The incomplete and fragmentary nature of metagenomic data presents challenges at every step of the typical workflow, which can be summarized into three stages: sampling, sequencing and annotation. The technical issues related with all of them influence the downstream analysis [70], making crucial the accurate tracking of sample metadata [60].

Metagenomic studies characterize both the composition of samples and the diversity across samples. One central problem in this context is to quantify the differences between environments and to correlate these differences with physical and biological properties. Finding optimal solutions to this problem has a wide potential impact in a range of environmental and medical applications. Classic approaches consider phylogeny of taxa found in different environments. For example, one reference is the UniFrac algorithm [52] and its variants, which currently represent the default

approach for comparing communities through high-throughput rRNA datasets. UniFrac uses a reference phylogenetic tree to define a distance between each pair of environments represented in the tree. Combined with a clustering algorithm or with principal component analysis, this method has achieved considerable results, allowing for example to discover that mammalian gut communities cluster primarily by diet and that the gut community is a highly distinct environment from other mammal microbiota [48].

In comparative metagenome analysis, the average genome size has served as a barcode to identify and compare metagenomes. A first attempt to estimate the AGS [70] is based on the calculation of the density of a set of marker genes that typically occur only once per genome. The prediction of the AGS directly from raw shotgun sequencing data establishes a relationship between genome size and environment, suggesting a clear correlation between environmental complexity and the diversity of the cellular repertoire that is required to cope with various external challenges. A number of recent publications [4, 41, 59] have described methods for estimating the AGS and have demonstrated substantial variations among communities, that reflect different geographical locations or depths in marine environments and also metabolic lifestyle [85]. From an evolutionary perspective, it is associated with genetic drift in small populations and genome streamlining in large one [46, 26].

In light of the central role of the average microbial genome size as an ecological parameter, we propose in this chapter an analytical argument to derive it. Our method is based on the interplay between genomic regularities in the form of scaling laws and the modeling of metagenomes. We will derive a metagenomic invariant that gives access to the moments of the probability distribution of genome sizes in a metagenome, the 0-moment being the number of organisms and the first moment being the AGS. The test of our prediction with simulated metagenomes gives satisfying results that we will further apply in empirical metagenomes.

4.2 Methods

All sections that contribute to the chapter refer to family scaling laws. To calculate them we used a set of "reference genomes" and their annotations in terms of domain family. We considered 981 bacterial species excluding from the data set all strains, as their presence would bias the family abundance profiles. The domain compositions of all analyzed bacterial genomes and the family annotations have been retrieved from the Pfam database release 27.0, specifically from the manually curated Pfam-A classification of proteins, a total of 8675 Pfam families appear in the reference bacterial genomes. Using the same procedure described in section 2.2.1, we verified that also Pfam families have scaling laws with family specific exponents and the relationship with the functional categories scaling is consistent with the previous findings.

The theoretical predictions about the functional form of the rescaled family abundance in metagenomes will be tested against a set of simulated metagenomes. To build them, we used the ensemble of 3568 bacteria provided by PFAM in the release 28.0. Compared to the release 27.0 from which the reference genomes have been obtained, 1445 new families have been added and 46

families have been killed.

The last section of the chapter investigates empirical metagenomes. Their functional and taxonomical annotations have been retrieved from the EBI database (Updated to 20/01/2016) employing the RESTful URLs to programmatically download the files of five Human Gut Microbiome projects, here addressed by their ID. A total of 248 metagenomes have been obtained, distributed between the projects as follows: 146 from ERP002469, 53 from ERP001956, 19 from SRP002423, 18 from SRP000319, and 12 from ERP001038. The files downloaded are the output results of the EBI pipeline 1.0 described in Ref.[35]. From the functional annotation files, only the matches of Pfam are retained. Among these, certain showed mutual overlapping. Hence, to avoid a wrong count of domains, the overlapping sequences are removed, keeping the matches with the highest score. The Pfam score selection is automatically applied by the EBI pipeline, therefore no further score cutoff has been applied. Since the pipeline 1.0 uses the release 24.0 of Pfam, we manually updated the family assignment to the release 28.0 by removing killed families and eventually merging them with others, following the evolution of Pfam family assignments over the years. While new families that descend from killed ones are forwarded, the ones newly created after the release 24.0 are absent. 5 families among the ones selected to calculate the moments of the genome size distribution in a metagenome are lost because of this. Specifically, 3 families with exponent $\beta \simeq 1$ (PF13365, PF13419, PF12704), and two with exponent $\beta \simeq 2$ (PF12681, PF12802). The total number of 16S sequences found in each metagenome is extracted from the files of taxonomic analysis, including also archaea and unassigned sequences. Archaea sequences constitute usually less than 1% of the total, reaching the 2.7% only in one sample. On the other hand, the percentage of unassigned 16S sequences is higher, reaching respectively an average of 16% and 8% in projects SRP002423 and ERP001038, although staying below 6% in every other metagenome. In addition, the number of 16S sequences belonging to Firmicutes and Bacteroidetes are calculated grouping the counts by phylum assignments and their relative abundances are calculated dividing by the total number of phylum-assigned sequences.

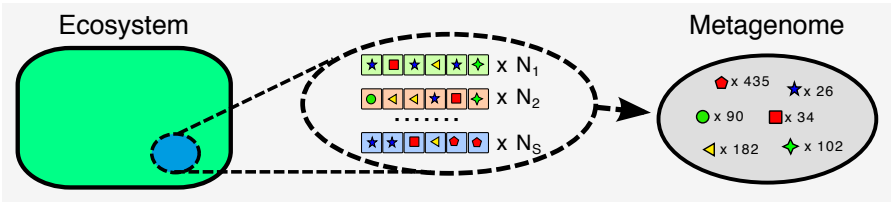

Figure 4.1: Illustration of a metagenome as a sample of environmental DNA. Metagenomics is defined as the analysis of genomes contained within an environmental sample [78]. After collecting the sample, the next step is sequencing the filtered DNA fragments and assigning protein annotations (different shapes in the figure) as well as functional annotation (colors of symbols). The DNA analyzed comes directly from the uncultured microbial community and belongs to different species (N_1 to N_S), but the assignment of DNA fragments to the correct taxa is one of the major challenges in metagenomics. In light of family-specific regularities in single genomes, we will express the the abundance of a protein family in the metagenome as a linear combination of the family abundances in single genomes.

4.3 The analytical implementation of family scaling laws results in the definition of a metagenome invariant.

In this section we will define the abundance of domain families in a metagenome $a^{MG}(f,m)$ integrating the known scaling laws that affect the abundance of domain families across single genomes (see Chapter 2). The family abundance $a^{MG}(f,m)$ rescaled by the pre-factor of the family-specific scaling laws and by the total number of organisms in the metagenome, should, on average, be solely a function of family exponent. As a consequence of that, we predict the existence of genomics invariants in metagenomics that allows the estimate of the moments of the genome size distribution (up to the second one).

Before starting the calculation about the family abundance in metagenomes, we will briefly discuss some useful properties of family scaling laws that were not presented in Chapter 2.

As mentioned in section 2.3, family scaling laws describe a mean behavior. The following equation holds :

$$a^G(f,g) = A_f n_g^{\beta_f} + \bar{\delta}_{f,g} \tag{4.1}$$

where $a^G(f,g)$ indicates the number of domains assigned to family f in genome g, n_g is the total number of domains in the genomes, the parameters A_f and β_f are family-specific and $\bar{\delta}_{f,g}$ is the family-specific fluctuation term. Ref. [31] shows that the fluctuations $\bar{\delta}_{f,g}$ among genomes with similar sizes follow a distribution with zero mean and a variance proportional to the average size $A_f n_g^{\beta_f}$, with a family-specific proportionality constant $\exp(Q_f)$. The observable Q_f is as an order parameter that measures the deviation from the Poisson behavior of the cross-species abundance distribution of family f, in particular it evaluates the mean to variance ratio of the abundance distribution. Q_f is defined as:

$$Q_f = \sum_b Q_{f,b} w_{f,b}.$$

The sum runs over the bins b of genome grouped by size and the elements in the sum are $Q_{f,b}$, which evaluates the deviation from Poisson behavior of the abundance distribution, and the sampling weight $w_{f,b}$. They are defined respectively as:

$$Q_{f,b} = (1 - \delta_{\mathrm{Var}_b(a_f),0}) \ \log \frac{\langle a_f \rangle_b}{\mathrm{Var}_b(a_f)} + \delta_{\mathrm{Var}_b(a_f),0} \left(\max_b \log \frac{\langle a_f \rangle_b}{\mathrm{Var}_b(a_f)} \right)$$

$$w_{f,b} = \frac{n_b n_b^+}{\sum_b n_b^2}$$

where δ_k, l is the Kroeneker's delta, a_f is the family abundance, n_b is the number of genomes in the bin b and n_b^+ is the number of non-zero entries of the family in the bin. It is noteworthy that Q_f exhibits a roughly linear anti-correlation with the scaling exponent β_f, meaning that the abundances that grow more with the genome size have larger fluctuations (see Fig. 4.2).

Since the family abundances are necessarily smaller than the the genome size we expect that for exponents β_f larger than 1, the prefactor A_f must be smaller than 1. Figure 4.3 shows that

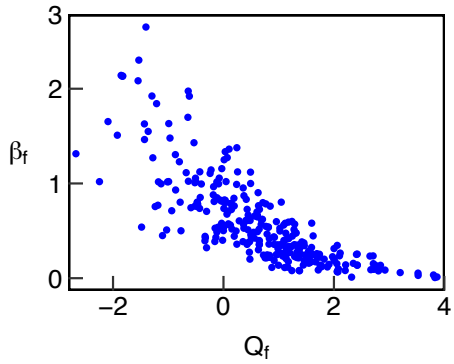

Figure 4.2: Higher scaling families have broader abundance distributions. The scaling exponent β_f anti-correlates with the order parameter Q_f (Spearman correlation is −0.85). To assure that scaling exponents are meaningful, only families with a Pearson correlation coefficient higher than 0.4 are displayed in this figure (the coefficient is calculated between the log of the genome size and the log of the family abundance with the same procedure described in 2.2.1.

in fact they are anti-correlated and, more precisely, the prefactor decreases exponentially with the exponent. An interesting consequence is that the prefactor can be expressed as a function of the exponent $A_f = k^{-\beta_f}$, with k independent on f. Finally, the only dependence of $a^G(f,g)$ on the family f is through β_f

$$a^G(f,g) = (n/k)^{\beta_f} \tag{4.2}$$

The interpretation of this result is that scaling laws for domain families actually revolve around a pivot point and evaluating scaling laws in a reference system centered in this pivot point may disentangle the interdependence between A_f and β_f. The estimated value for the constant k is of the order of 10^3.

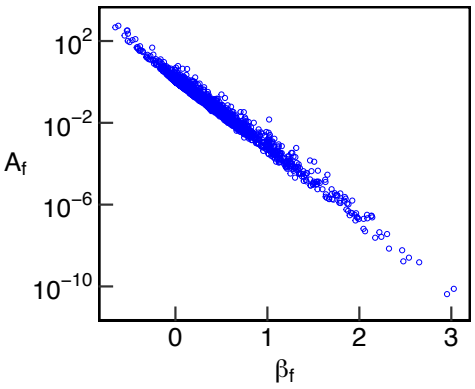

Figure 4.3: The scaling exponent of domain families is anti-correlated with the prefactor. This plot shows the prefactor A_f (y-axis) as a function of the scaling exponent β_f (x-axis) and the lin-log scales highlights that A_f decreases exponentially. Each dot corresponds to a Pfam family, only families with Pearson correlation coefficient higher than 0.4 are retained (see section 2.2.1 for the details).

4.3.1 Analytical derivation of the abundance of a protein family in a metagenome

This section will use the regularities of genome composition to analytically study the abundance of a domain family in a metagenome. The first point is finding a convenient way to express $a^{MG}(f,m)$ in terms of the abundances of domain families in single genomes. The development of the resulting expression leads to the discovery of a new family-invariant observable closely related to

the size distribution of constituent genomes. A metagenomic sample is composed by an ensemble of N_{tot} microbial genomes belonging to S different species. The taxonomic composition of a metagenome is in principle not known, this is due to the sampling procedure and the subsequent DNA fragmentation necessary to sequence and annotate the sample. Assuming that the S species are represented with abundances $N_1, \ldots, N_S$ (with $\sum_i^S N_i = N_{tot}$), the abundance of a generic family f in a metagenome m can be written as:

$$a^{MG}(f,m) = \sum_{g=1}^{S} a^{G}(f,g) N_g \tag{4.3}$$

This equation requires the following implicit hypothesis to be true:

(i) organisms of the same species have identical genomes;

(ii) the family f is present in each genome contained in the metagenome;

(iii) no gene subsampling took place.

Equation 4.3 combined with eq. 4.1 about the scaling of domain families, defines the abundance of a scaling family inside a metagenome as

$$a^{MG}(f,m) = \sum_{g=1}^{S} (A_f n_g^{\beta_f} + \bar{\delta}_{f,g}) N_g$$

The first term of the sum depends on the genome g only by the size n_g, so the terms of the sums can be grouped by species genome size

$$a^{MG}(f,m) = \sum_{n=1}^{\infty} A_f n^{\beta_f} \sum_{g,n_g=n} N_g + \sum_{n=1}^{\infty} \sum_{g,n_g=n} \bar{\delta}_{f,g} N_g$$

After dividing by A_f and N_{tot}, the previous equation becomes:

$$\Gamma(f,m) = G_m(\beta_f) + \Delta(f,m) \tag{4.4}$$

where we defined three new quantities as:

$$\Gamma(f,m) \equiv \frac{a^{MG}(f,m)}{N_{tot} A_f} \tag{4.5}$$

$$G_m(\beta_f) \equiv \sum_{n=1}^{\infty} n^{\beta_f} P_m(n) \tag{4.6}$$

$$\Delta(f,m) \equiv \frac{1}{A_f} \sum_{n=1}^{\infty} \sum_{g,n_g=n} \bar{\delta}_{f,g} \tilde{P}_m(g) \tag{4.7}$$

where $\tilde{P}_m(g) \equiv N_g/N_{tot}$ is the distribution of species and $P_m(n) \equiv \sum_{g,n_g=n} \tilde{P}_m(g)$ is the distribution of genome sizes inside the metagenome m.
We will now examine in detail the properties of these three functions.

$\Gamma(f, m)$. This quantity depends on the total number of genomes N_{tot} that constitutes the metagenome. N_{tot} is a priori not known, therefore in case of empirical metagenome we will use the rescaled version $N_{tot}\Gamma(f, m)$ and simply refer to it as the *rescaled family abundance*. If hypothesis *(ii)* in eq. 4.3 is violated, that is the family f is present in $N^\star < N_{tot}$ genomes, then the rescaled family abundance will be underestimated.

$G_m(\beta_f)$. The functional form of $G_m(\beta)$ is determined by the genome size composition of the metagenome and this fact makes $G_m(\beta)$ a good candidate as a metagenomic barcode. Moreover, the transformed size distribution $G_m(\beta_f)$ is a family-invariant, since its dependence on the family f is reduced to the value of its scaling exponent β. Hence, when the term of noise $\Delta(f, m)$ is negligible, we expect the rescaled family abundances $\Gamma(f, m)$ to follow the well-defined functional form of $G_m(\beta)$, thus giving access to the information about the genome size distribution.

$\Delta(f, m)$. The fluctuations of the family abundance depend on $\tilde{P}_m(g)$, the distribution of species mixed in the metagenome, since the fluctuations $\bar{\delta}_{f,g}$ are genome-specific. Hence, we cannot affirm anything general about $\Delta(f, m)$, since it depends on the specific composition of each metagenome. We can however examine two particular cases of metagenome composition and their effect on the fluctuations.

(i) Uniform species sampling. Assuming that the distribution of species composing a metagenome samples uniformly the space of genomes with similar sizes n, the fluctuations should cancel out, since we know that the fluctuations $\delta_{f,m}$ have zero mean value among reference genomes with similar sizes. In this case Eq.4.4 reduces once again to the equivalence between $\Gamma(f, m)$ and $G_m(\beta_f)$. This hypothesis models a metagenome taken from an environment with high diversity.

(ii) Hypothesis of single species sampling. We now make the opposite assumption: a metagenome composed solely by genomes of a single species g with size n_g, i.e. a delta-like species distributions $\tilde{P}_m(g) = \delta_{g,g'}\delta_{n,n_g}$. In this case the fluctuations sum up constructively and the term of noise becomes

$$\Delta(f, m) = \frac{\bar{\delta}_{g,f}}{A_f}$$

This deviation is genome-specific and is not necessarily large. In the worst cases, anyway, $\bar{\delta}_{g,f}$ can be of the order of $(e^{-Q_f}A_f n_g^{\beta_f})^{1/2}$ and consequentially

$$\Delta^{max}(f, m) \sim \left(\frac{n_g^{\beta_f}}{e^{Q_f}A_f}\right)^{1/2}$$

Since both Q_f and A_f are decreasing functions of β (see section 4.3), we expect this fluctuation to increase with the value of the exponent.
This hypothesis models metagenomes with very low diversity, such as those taken from a bacterial colony.

Equation 4.4 is valid in the hypothesis that metagenomes are not subject to gene subsampling. This is clearly an approximation that allow us to derive in a few steps the metagenome invariant $G_m(\beta)$. To test the robustness of eq. 4.4, we now discuss three different scenarios, that presume the existence of biases during the sequencing and annotation of metagenomes.

A first possible bias on eq. 4.4 is due to removal of identical sequences. In certain analysis pipelines, almost identical sequences are identified and copies are removed from the sample prior to being annotated. The extent of this subsampling depends on the particular pipeline: if no assembly of randomly fragmented sequences into genes is performed, it is unlikely that two small sequences result identical. If instead identical genes are removed, this implies that N_g identical genomes are counted as 1. This fact changes the composition of the metagenome at a species level, changing $\tilde{P}_m(g)$. However, it does not affect the calculations of $\Gamma(f, m)$ but only the form of $G_m(\beta)$. The sequencing of environmental samples is particularly difficult [60, 86], because of the fragmented and partial nature of the protein coding sequences. Therefore it is legitimate to assume that a fraction of the original sequences may get lost. We can model this by saying that each domain among the n_m contained in the mix of genomes is kept with probability p. This leads to a binomial-distributed metagenome size k

$$f(k; n_m, p) = \Pr(X = k) = \binom{n_m}{k} p^k (1-p)^{n_m - k} \tag{4.8}$$

Each domain among the k retained might belong to the family f with probability $p_f = a^{MG}(f, m)/n_m$. Therefore, the k domains are partitioned in the F families with multinomial probability

$$g(a_1, \ldots, a_F; k, p_1, \ldots, p_F) = \frac{k!}{a_1! \cdots a_F!} p_1^{a_1} \cdots p_F^{a_F}$$

Assuming that the probability p is small and n_m is large, it is possible to approximate the binomial of eq. 4.8 with a Poissonian distribution with average pn_m. Under this assumption, the joint probability becomes

$$h(k, a_1, \ldots, a_F) = \frac{(pn_m)^k e^{-pn_m}}{k!} \frac{k!}{a_1! \cdots a_F!} p_1^{a_1} \cdots p_F^{a_F}$$

Moreover, by exploiting the fact that $k = \sum_i a_i$, $p_f = a^{MG}(f, m)/n_m$ and $n_m = \sum_i a^{MG}(f, m)$, the joint probability can be factorized in

$$h(k, a_1, \ldots, a_F) = \prod_i^F \frac{(pa^{MG}(f, m))^{a_i}}{a_i!} e^{-pa^{MG}(f,m)}$$

implying that each family abundance is sampled independently, following a Poissonian distribution with mean $pa^{MG}(f, m)$. Thus each sampled family abundance a_f is, on average, a fraction p of the original abundance $a^{MG}(f, m)$. The effective rescaled family abundance is then

$$\tilde{\Gamma}(f, m) \simeq \frac{pa^{MG}(f, m)}{A_f N_{tot}} \tag{4.9}$$

As a consequence, when plotted in lin-log scale, $\tilde{\Gamma}(f, m)$ is only a vertical translation of the original rescaled abundance $\Gamma(f, m)$, thus maintaining its functional form.

In the typical metagenome analysis workflow, the annotation procedure follows the sequencing of DNA fragments. This step presents major computational challenges and is inevitably a source of bias. Annotation begins with the identification of features of interest, like genes, and continues by assigning to these features functional and taxonomical units based on homology searches against available data [78]. The matches are associated with a score representing their likelihood of not having been emitted by chance. Matches with a score too low are usually rejected. The score depends also on the length of analyzed sequences and shorter ones are more likely to have a low score. Since the protein coding sequences in metagenomic data sets tend to be fragmentary, it is quite likely that many of the sequences coding for these proteins are too short to make significant matches. Moreover, the length of domain coding sequences can vary between different families, thus leading to their differential subsampling caused by fragmentary sequences. This implies that the effective rescaled family abundance can diversely underestimate the original value depending on the considered family.

4.3.2 The metagenomic invariant gives access to the moment of the distribution of genomes size in the metagenome

Under the assumption that the fluctuation term $\Delta(f, m)$ is negligible, eq. 4.4 establishes a direct relation between the term $G_m(\beta)$ and the metagenome invariant $\Gamma(f, m)$

$$\Gamma(f, m) = G_m(\beta_f). \tag{4.10}$$

Provided that there are domain families that scale with an integer exponent $\beta = k$, the function $G_m(\beta)$ gives exactly the k-th moment μ_k^m of the genome size distribution

$$G_m(\beta = k) = \sum_{n=1}^{\infty} n^k P_m(n) \equiv \mu_k^m \tag{4.11}$$

Since $\Gamma(f, m)$ can be extracted from empirical data, we have that eq. 4.10 estimates the moments μ_k^m of the probability distribution of genome sizes in the metagenome. The distribution $P_m(n)$ is usually unknown in sampled metagenones, but later on we will study two cases, where analytical calculations are feasible and derive the moments of the distribution.
In general, a metagenome is composed of S species with l different sizes n_i. According to eq. 4.11 the transformed genome size distribution $G_m(\beta)$ is a weighted sum of exponentials in β

$$G_m(\beta) = P_m(n_0) n_0^\beta + \cdots + P_m(n_l) n_l^\beta \tag{4.12}$$

Sorting n_i by increasing size and grouping n_0, we get

$$G_m(\beta) = n_0^\beta \left(P_m(n_0) + \sum_{i=1}^{l} \left(\frac{n_i}{n_0} \right)^\beta P_m(n_i) \right) \tag{4.13}$$

For β large enough, the coefficient reduces to $P_m(n_0)$ and $G_m(\beta)$ reduces to a single exponential in β. In principle it is possible to fit the tail of $\Gamma(f, m)$ to get $P_m(n_0)$. However the closer n_0 and n_i are, the slower $P_m(n_i)$ disappears and the range of empirical exponents β is limited between 0 and 3. For highly peaked distributions, approximating $G_m(\beta)$ with an exponential is correct even for small values of β.

To simplify the analytical calculations we now consider the genome size n as continuous, thus substituting the sums with integrals and the size distribution with a probability density function $p_m(n)$. Under this approximation, the transformed genome size distribution becomes

$$G_m(\beta) = \int_0^\infty n^\beta p_m(n) \mathrm{d}n. \tag{4.14}$$

which corresponds exactly to the Mellin's transform of $p_m(n)$.

We then consider two possible distributions whose Mellin's transform has a simple analytical form. As first example, we consider $p_m(n)$ equal to the uniform distribution of width Δn centered in $\bar{n}$

$$p_m(n) = \frac{1}{\Delta n} \left[\theta(n - (\bar{n} - \Delta n/2)) + \theta((\bar{n} + \Delta n/2) - n)\right]$$

Since the genomes size are necessarily positive, the relation $\bar{n} \geq \Delta n/2$ must be true. Under this condition the transform of $p_m(n)$ is

$$G_m(\beta) = \frac{\bar{n}^{\beta+1}}{\Delta n(\beta+1)} \left[\left(1 + \frac{\Delta n}{2\bar{n}}\right)^{\beta+1} - \left(1 - \frac{\Delta n}{2\bar{n}}\right)^{\beta+1}\right]$$

Through a series expansion of $(1+\epsilon)^\beta$ up to the fourth order in ϵ, the transformed size distribution can be approximated as

$$G_m(\beta) = \bar{n}^\beta \left[1 + \beta(\beta-1)\left(\frac{\Delta n}{2\bar{n}}\right)^2 + o(\Delta n/2\bar{n})^4\right] \tag{4.15}$$

which reduces to the transform of $p_m(n) = \delta(n - \bar{n})$ as $\Delta n \to 0$.

An other easily tractable case is the one of delta distribution. Assuming that the genome size distribution is particularly peaked, we can approximate $p_m(n)$ with the Dirac's delta centered in $\bar{n}$, whose transform $G_m(\beta)$ is the exponential in β with base $\bar{n}$

$$p_m(n) = \delta(n - \bar{n})$$

$$G_m(\beta) = \bar{n}^\beta \tag{4.16}$$

To sum up, equations 4.15 and 4.16 represent two possibilities for the analytical form of the metagenome invariant $G_m(\beta)$. In the following paragraphs, we present two different strategies to access the moments of the genome size distribution.

Selection of families for $\beta \simeq k$. Since eq. 4.11 states that the metagenome invariant $G_m(\beta)$ gives the $k-th$ moment when β is an integer value, it is possible to isolate scaling families with exponent close to 0, 1 or 2 and to estimate the number of genome N_{tot}, the average genome size $\bar{n}$ and the variance Var(n). With this method, one can obtain reliable values of the rescaled family abundance $\Gamma(f, m)$ with which approximate the theoretical value of $G_m(\beta)$.

Fit of an assumed transformed size distribution. When log-transformed and multiplied by N, eq. 4.15 and eq. 4.16 become respectively:

$$NG^2_{log}(\beta; \log N, \log \bar{n}, \epsilon) = \log N + \beta \log \bar{n} + \log(1 + \beta(\beta - 1)\epsilon^2) \tag{4.17}$$

$$NG^1_{log}(\beta; \log N, \log \bar{n}) = \log N + \beta \log \bar{n} \tag{4.18}$$

Bearing in mind that $NG_m(\beta)$ equals $N\Gamma(f, m)$ (eq. 4.10) and that $N\Gamma(f, m)$ is measurable, it is possible to fit the above equations with parameters N and $\bar{n}$.
To implement the fitting procedure, we first removed families with negative β. Then, the values of $\Gamma(f, m)$ are partitioned by β in bins of width $w = 0.01$ and centered in

$$x_l = \beta_{min} + (l + 0.5)w$$

and finally the logarithm of $N\Gamma(f, m)$ is fitted against the logarithms of the binned, averaged values. We repeated the fit for thirty different occurrence cutoffs, from 0.7 to 0.99, with steps of 0.01 and averaged the resulting parameters. In case of the first equation the width of the uniform distribution is obtained as $\Delta n = 2\epsilon\bar{n}$.

4.4 The mean genome size and the number of genomes in a metagenome are estimated reliably in simulated metagenomes.

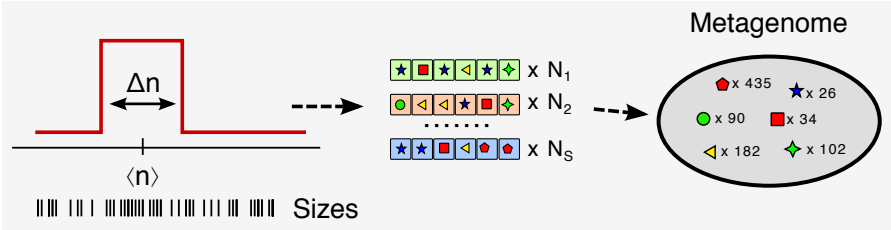

Figure 4.4: Illustration of procedure to simulate a metagenome. We fixed the distribution of genome size as the uniform distribution with width Δn and mean $\langle n \rangle$. We randomly extracted N_S sizes and matched them with genomes of the same size in our database. The selected genomes, along with their domain annotations, will build the simulated metagenome. Different symbols correspond to different domain families, their color represents the functional assignment.

This section tests the theoretical predictions made in section 4.3.2 one about the estimation of the moments of the genome size distribution in a metagenome. in order to test the effectiveness and reliability of this estimate, we created artificial metagenomes assembling genome annotation of randomly selected bacterial species. The first step to build a metagenome, is to fix the simulation parameters, that are the number of genomes that will build each metagenomes N_{true} and the genome size distribution $Q_m(n)$. A simple choice for $Q_m(n)$ is the uniform distribution with mean $\bar{n}$ and variance $\sigma^2 \simeq 12\Delta n^2$:

$$Q_m(n;\bar{n},\Delta n) = \frac{1}{\Delta n}\left[\theta\left(n-(\bar{n}-\Delta n/2)\right)+\theta\left((\bar{n}+\Delta n/2)-n\right)\right]$$

where $\theta(x-\bar{x})$ is the Heaviside step function. In this way we are able to calculate the Mellin's transform of $Q_m(n)$ and this allows analytical calculations of $G_m(\beta)$. For each metagenome we draw the mean size $\bar{n}$ from the interval $[n_{min}+\Delta n/2, n_{max}-\Delta n/2]$ with equal probability. Then N_{true} values of genome size are extracted with probability $Q_m(n;\bar{n},\sigma^2)$ and associated with a genome of equal size. In case of more than one genome with the picked size, one of these correspondences is chosen randomly with uniform probability. In order to pick different genomes uniformly, the N_{true} random sizes are extracted independently. We created 1500 metagenomes an used this set to test our predictions.

4.4.1 The rescaled family abundance in simulated metagenomes shows clear scaling with family exponent.

As already mentioned in sec. 4.3.1, the value of N in empirical metagenomes is not known and we can only calculate $N\Gamma(f,m)$ as $a^{MG}(f,m)/A_f$. For the sake of brevity, here we will refer to it simply as $\Gamma(f,m)$ as in lin-log scale the only difference between them is a constant vertical shift.
For all the simulated metagenomes $\Gamma(f,m)$ has a well defined functional form in the scaling exponent β (see Fig. 4.5). The exponential-like behavior is not an artifact due to the interdependence between the prefactor A_f and the scaling exponent β_f (eq. 4.2), since $1/A_f$ does not depend on the metagenome. In the next paragraph, other observations will confirm that the dependency of $\Gamma(f,m)$ on β is genuinely shaped by the family abundance $a^{MG}(f,m)$.

In order to obtain a reliable rescaled family abundance $\Gamma(f,m)$ that will serve to obtain the transformed size distribution $G_m(\beta)$, it is necessary to select carefully the data to keep. Figure 4.5 shows that the rescaled family abundance underestimates the theoretical value of $G_m(\beta_f)$ when the family f has a low occurrence, i.e. when it is absent in a large fraction of genomes composing the metagenome m, as predicted in sec. 4.3.1. Since we cannot know the true occurrence of f in empirical metagenomes, it is not possible to use it as a selecting parameter for the families that better align with $G_m(\beta)$. A good strategy, however, is to impose a threshold on the minimum occurrence of f among reference genomes. This choice causes an effective cutoff on the true occurrence, since it selects families that are more likely to appear in a generic bacterial genome. The exponent β_f and the occurrence in reference genomes are completely uncorrelated and this guarantees that imposing a general cutoff of occurrence will maintain the same proportions of families having different exponents.

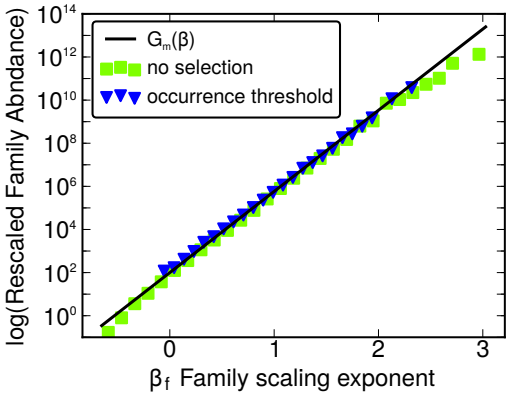

Figure 4.5: The rescaled family abundances scales clearly with the family exponent, although a low family occurrence causes the underestimation of the transformed size distribution. Binned log-values of rescaled family abundances $N\Gamma(f,m)$ (y-axis) are plotted as function of the family scaling exponent β_f (x-axis). All plotted data describe a well defined straight line, exhibiting an exponential-like behavior. When no selection is applied (green squares), the values underestimate the theoretical transformed distribution $G_m(\beta_f)$ (black line) because of the low occurrence of some families in the metagenome. Considering only families that appear in more than 80% of the reference genomes (blue triangles) we obtain an effective cutoff on the true occurrence in the metagenome and the theoretical behavior is found. Metagenome info: N = 100, $\langle n\rangle$ = 5819 , σ = 102

Since the rescaled family abundances $\Gamma(f,m)$ should describe the transformed size distribution $G_m(\beta)$, its plot in β should reflect the different composition of metagenomes. Here we confirm this fact, observing the differences between metagenomes with diverse average size and variance.
Considering similarly disperse distributions of sizes with different mean values $\langle n\rangle$, the scaling of the rescaled family abundance $\Gamma(f,m)$ changes visibly (Fig.4.6). In fact, the average genome size is the base of the exponential that constitutes $G_m(\beta)$ and in lin-log scale this translates into two straight lines with well distinct slopes.

On the contrary, when almost identical values of $\langle n\rangle$ and different values of σ are considered, size distributions correspond to barely discernible $\Gamma(f,m)$. Notably, we would expect a deviation around $\beta \sim 2$, where the variance adds to $\langle n\rangle^2$ giving $\langle n^2\rangle \sim \Gamma(f,m|\beta_f = 2)$. Unfortunately the variances are, in most cases, of the same order of $\Gamma(f,m|\beta_f = 2)$ fluctuations or even smaller. In addition, the logarithmic scale of the y-axis tends to hide differences between quantities if they are of smaller orders of magnitude. In general, the variations of $\Gamma(f,m)$ are caused by fluctuations of $a^{MG}(f,m)$ and are therefore in units of $1/A_f$ (fig.4.3). The prefactor scales exponentially in β and the variations of $\Gamma(f,m)$ inherit the same behavior.

4.4.2 The total number of sampled genomes can be estimated reliably in simulated metagenomes

We carried out the calculation with both methods described in sec. 4.3.2. First we present the results obtained through the selection of families with 0 scaling exponent and then fitting the function

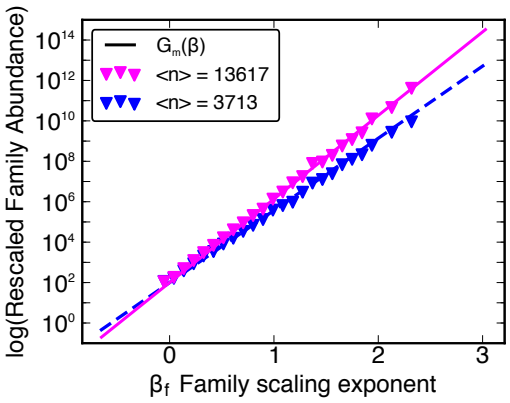

Figure 4.6: Two metagenomes with different average genome size $\langle n \rangle$ are easily discernible by the scaling of their rescaled family abundances. Considering two metagenomes with almost identical σ and different $\langle n \rangle$, $\Gamma(f, m)$ follows well distinct curves. The lines are the theoretical transformed distribution $G_m(\beta_f)$ for the two metagenomes. Metagenomes info: $N = 100$, $\langle n \rangle = 13617$, $\sigma = 101$ (pink triangles and solid line) and $N = 100$, $\langle n \rangle = 3713$, $\sigma = 101$ (blue triangles and dashed line). Occurrence threshold: 0.8.

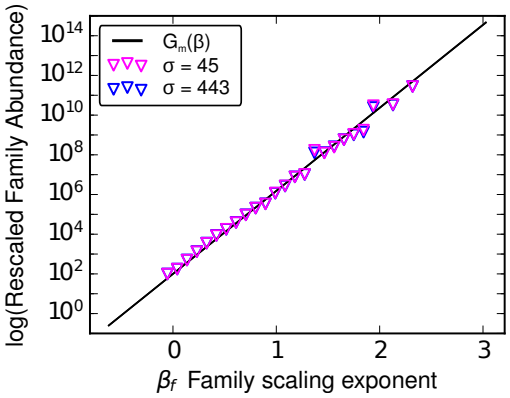

Figure 4.7: The rescaled abundances of two metagenomes with identical $\langle n \rangle$ and very different σ are hardly discernible. The rescaled abundances (y-axis) binned and averaged by scaling exponent (x-axis) for two metagenomes with different variance of sizes σ^2: $\sigma = 45$ (pink triangles) and $\sigma = 443$ (blue triangles). The solid black lines are the theoretical transformed distribution $G_m(\beta_f)$ for the two metagenomes and clearly coincide. Metagenomes info: $N = 100$, $\langle n \rangle = 15135$, $\sigma = 45$ and $N = 100$, $\langle n \rangle = 15329$, $\sigma = 443$. Occurrence threshold: 0.8.

$G_m(\beta)$. In both cases the simulations match with the predicted values.

Under our assumption eq. 4.10 holds, which written extensively is:

$$\frac{a^{MG}(f,m)}{A_f\, N} = \sum_{n=1}^{\infty} n^{\beta_f} P_m(n). \tag{4.19}$$

Because of the normalization of the size distribution, this equation evaluated in $\beta = 0$ gives access to the number of genomes N, hence we selected families with reliable scaling law, and whose exponent is close to zero.

The high number of families with $\beta \simeq 0$ allows us to be highly selective in picking them. Only the families with exponent values in $[-10^{-3}, 10^{-3}]$ are considered and the ones with occurrence lower than 0.99 among reference genomes are rejected. Then, only those that have a "goodness of fit" index s^f_{LS} higher than 0.999 are kept (for the details of the calculation of the parameter s^f_{LS} see section 2.2.1 of the thesis). 13 families overcome the selection, noticeably they all have an average abundance among genomes extremely close to 1. For this reason, the prefactors A_f of the power laws are manually set to 1, avoiding the propagation of the existing, although small, errors.
For each selected family, we calculate the empirical value of $N\Gamma(f, m|\beta_f = 0)$ that in eq. 4.19 gives

the number of genomes. The final estimation of N is the averaged value.
In general, the rescaled family abundance $\Gamma(f, m)$ of selected single-copy families gives an incredibly good estimate of the number of genomes in the metagenome.

The strategy of obtaining N by fitting the exponential $G_m(\beta)$ (precisely eq. 4.18) reveals to be good, although it gives worse results than the manual selection of families with $\beta \simeq 0$. The estimates of the number of genomes generally improves as the occurrence threshold grows (fig. 4.8B). This depends on the fact that reducing the incidence of low occurring families make $\Gamma(f, m)$ more similar to $G_m(\beta)$ (fig.4.5). The errors associated with each occurrence threshold are calculated as the mean squared error of the estimated N among the metagenomes

$$\mathrm{MSE}_{occ} = \frac{1}{M}\sum_{m=0}^{M}(NG^i_{log}(0; \mathrm{p}^i_{fit,m}) - N_m)^2$$

where $\mathrm{p}^i_{fit,m}$ are the parameters of the i-th function NG^i_{log}, fitted with the $\Gamma(f, m)$ filtered by occurrence and M is the number of simulated metagenomes.

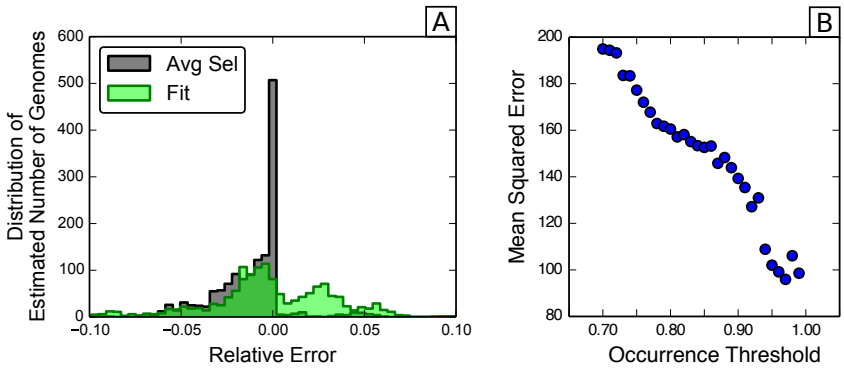

Figure 4.8: (A) Both the selection of families with $\beta_f = 0$ and the fit give a very good estimate of the number of genomes. Distributions of relative errors of N for the fit (green) and the average abundance of single-copy families (black).**(B) The error in the estimate of the number of genomes given by fit decreases as the occurrence threshold increases.** Mean squared error of the estimate (y-axis) as a function of the occurrence threshold (x-axis). Simulations info: 1500 metagenomes, $N = 1000$, $\sigma = 100$.

4.4.3 The average genome size can be estimated reliably in simulated metagenomes.

We carried out the calculation with both methods described in sec. 4.3.2. First we present the results obtained through the selection of families with 1 scaling exponent and then fitting the function $G_m(\beta)$. In both cases the simulations match with the predicted values.

The lower number of families with $\beta \simeq 1$ imposes to be less selective than the previous case. Firstly only the families with value of β in $[0.95, 1.05]$ are considered. Then, a lower cutoff on occurrence is imposed, removing every family that occurs in less than 90% of reference genomes.

With these constraints, only 12 families are retrieved. The estimate of the average genome size associated with a selected linear family f is calculated as

$$\langle n \rangle_f = \frac{a^{MG}(f,m)}{A_f N}$$

where N is the total number of genomes calculated selecting families with $\beta = 0$. Selected families have rescaled abundances $\Gamma(f,m)$ that fluctuate with different intensities around the true average genome size but their values averaged give a good estimate of $\langle n \rangle$ (fig. 4.9A).

The estimates of the average genome size obtained by fitting the exponential equation 4.18) describe well the true values (figs.4.9A). For a large range of occurrence thresholds, the deviations from the theoretical behavior remain the same but increase rapidly as the cutoff surpasses 0.95 (figs.4.9B), i.e. when the maximum exponent of the retained families decreases due to the scarce number of families with high β. Since the average genome size $\langle n \rangle$ corresponds approximatively to the slope of the straight line described by NG^i_{log} vs. β in lin-log scale, its value is easy to measure just as long as the range of β is large. The deviations are calculated as

$$\mathrm{MSE}_{occ} = \frac{1}{M}\sum_{m=0}^{M}(NG^i_{log}(1; \mathrm{p}^i_{fit,m}) - \langle n \rangle)^2_m$$

where $\mathrm{p}^i_{fit,m}$ are the parameters of the i-th function NG^i_{log}, fitted with the rescaled family abundances $\Gamma(f,m)$ filtered by occurrence and M is the number of simulated metagenomes.

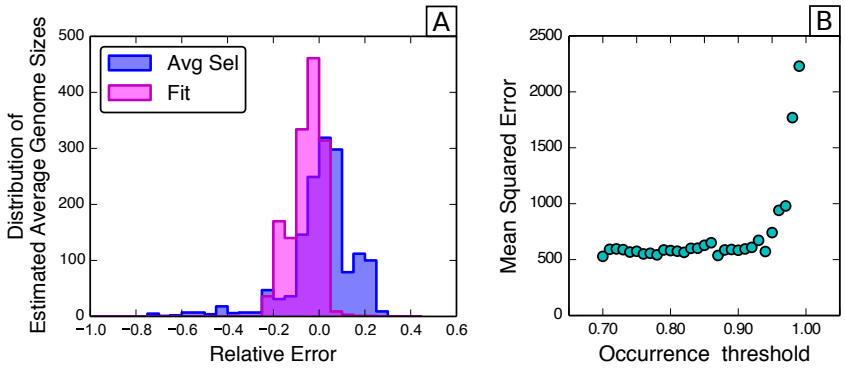

Figure 4.9: (A) Both the selection of linear families ($\beta \sim 1$) and the fit give a very good estimate of the average genome size. Distributions of relative errors in the estimate of the average genome size obtained by the fit (magenta) and averaging the rescaled abundances of selected families (blue). **(B) The mean errors in the estimate of the average genome size are are almost constant for a wide range of occurrence thresholds but increase rapidly for values** > 0.95 Simulations info: 1500 metagenomes, $N = 1000$, $\sigma = 100$.

4.4.4 The variance of the genome size distribution deviates from the predicted behavior.

Again, both methods described in sec. 4.3.2 will be applied to estimate the variance of the genome size distribution. Neither selecting families with exponent equal to 2 nor the fit of $G_m(\beta)$ give

satisfying results and we are forced to conclude that the variance can not be reliably estimated.

The low number of families with $\beta \simeq 2$ imposes particularly loose selection conditions. Applying an occurrence cutoff for of 0.5 and considering only families with $\beta \in [1.9, 2.1]$, only 10 families are retained. The estimates of the variance are calculated as

$$\mathrm{Var}_f = \frac{a^{MG}(f,m)}{A_f N_{true}} - \langle n \rangle^2_{true}$$

where N_{true} and $< n >_{true}$ are respectively the true number of genomes mixed in the metagenome and the true average genome size. When the scaling exponent β_f is exactly 2, the rescaled family abundance $\Gamma(f,m)$ follows qualitatively the theoretical behavior of the transformed size distribution $G_m(2)$ predicting the value $\langle n^2 \rangle$ and thus the variance $\mathrm{Var}(n)$, although showing large fluctuations. However, as the distance of the exponent β_f from 2 grows, the estimated values diverge from the true value (fig.4.10). It is important to notice that for exponents close to 2, $1/A_f$ has values of order

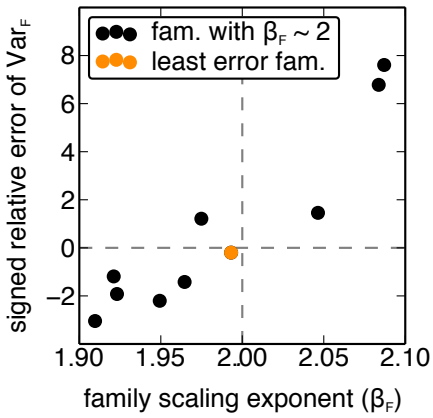

Figure 4.10: Only one selected family gives a fairly good estimate of the variance, while the others strongly correlate with the distance from 2 of their exponent. The signed relative error of the estimated variance related to the selected families (y-axis) averaged over all the simulated metagenomes, shows a clear correlation with the family exponent (x-axis). Among the 10 selected families (black dots), the one with exponent closer to 2 (orange dot) exhibits a particularly small signed error. Simulations info: 1500 metagenomes, $N = 1000$, $\sigma = 3000$.

$10^6 - 10^7$ (fig 4.3). Hence deviations of $a^G(f,g)$ imply fluctuations at least of the same magnitude in $\langle n^2 \rangle$. This fact has a drastic effect on the precision with which the variance can be calculated. Whenever the variance is of smaller order of magnitude than 10^7, it is dominated by the noise and its true value remains impossible to determinate. For this reason, we will not apply our theoretical prediction to the calculation of the variance in real metagenomes.

The fit of the transformed uniform distribution fails badly in estimating the value of the variance, even for high values of $\mathrm{Var}(n)$, that according to what we discussed above, could be over the noise threshold. One possible reason to explain it could lay in the use of NG^2_{log}, which is an

approximation of the original transformed size distribution. This choice, however, is done in order to make the fit feasible and thus cannot be changed.

4.5 The mean genome size and the number of genomes are estimated reliably in real metagenomes.

The calculation on simulated metagenomes made in sec. 4.4 confirmed the theoretical predictions about the number of genomes N_{tot} and the average genome size $\langle n \rangle$ in a metagenome. This section tests our methods in real metagenomes. The first check involves the rescaled family abundances $\Gamma(f, m)$, which in empirical metagenomes exhibits the same, well-defined, exponential dependency in β observed in simulated metagenomes. After verifying that $\Gamma(f, m)$ follows the predicted be-

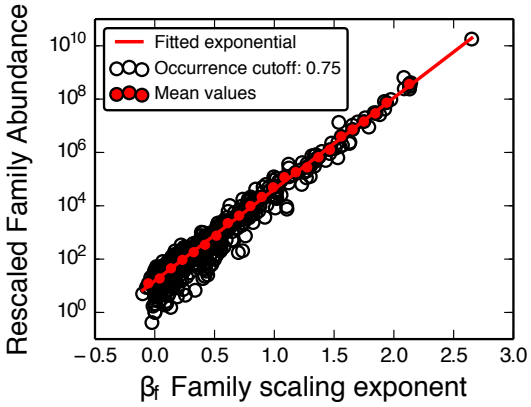

Figure 4.11: The rescaled family abundance exhibits a perfectly exponential dependency in β also in empirical metagenomes. The rescaled family abundance (y-scale) plotted in lin-log scale as a function of β (black dots). Data are from the metagenomics sample (sample id-code ERR056990) collected from the gut microbiota of three-month-old infants [74]. The mean values per bin of β (red dots) are perfectly fitted by an exponential (red line).

havior, the goal is to estimate the total number of genomes and the average genome size of the empirical metagenomes. As explained in section 4.3.2, two different methods proved to be efficient: the selection of families with integer scaling exponent and the fit of the theoretical $G_m(\beta)$.

However, since we are dealing with real metagenomes, it is important to consider if they may be affected by biases happened at the stage of sequencing. Section 4.3.1 explained how the sampling of coding sequences might take place during the sequencing process, thus leading to a change in the family abundances. The present paragraph shows that sampling probably took place in considered empirical metagenomes, leaving the family proportions intact. This fact will then be used to easily extend the calculations of the rescaled family abundance to the present case. The EBI pipeline v1.0 applies quality controls on analyzed sequences, removing sequences that are too short or have a low complexity and clustering identical copies [35]. Throughout this process it is thus high likely

that a sampling of families took place. An evidence of the sampling appears comparing the number of 16S sequences and the abundances of the 13 selected single-copy families (scaling exponent $\beta \simeq 0$) in a metagenome. These two quantities appear to be proportional (Fig.4.12), with 16S matches that are constantly more than single-families by an average factor of ~ 10 across projects. Since 16S genes in complete bacterial genomes appear in a low number of copies, constant within species [83], the number of 16S sequences in a metagenome should be proportional to the total number of genomes contained, while the abundance of single-copy families should correspond exactly to this number. However, the number of 16S copies per genome is usually between 1 and 7 [83] and thus it is not compatible with the observed disproportion. This fact suggests that 16S and the 13 families must have been sampled and that the process acted differently on the two kind of sequences. The differentiation might rise from the annotation process, which is done with two different tools for 16S and other sequences [35].

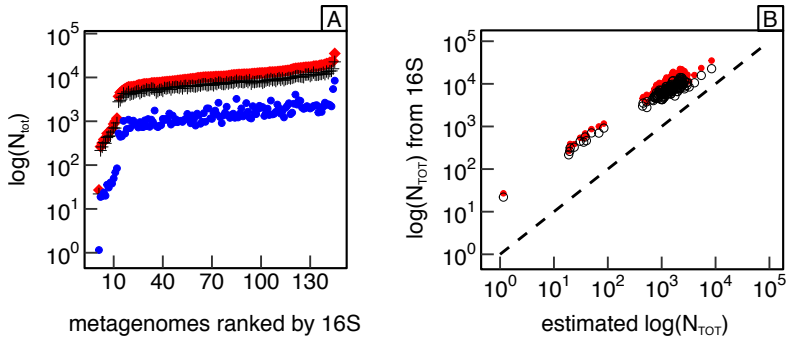

Figure 4.12: The number of genomes estimated through 16S analysis is systematically higher than the value obtained through the selection of families with $\beta \sim 0$. (A) This plot displays the comparison between the number of genomes in the metagenome N_{tot} estimated through the selection of families with $\beta \sim 0$ (blue dots) and the 16S annotation counting all different taxa (red diamonds) or just bacterial species (black crosses). The values of N available through EBI are one order of magnitude higher that our estimation, suggesting the presence of uniform sampling for families. (B) Scatterplot of the estimated number of genomes using 16S annotations (y-axis) versus the total number of genomes estimated using a set of constant-scaling families. Red dots are the estimations made using all taxa, black circles with only bacteria and the dashed line represents the bisect. Data displayed are gut metagenomes of european women [39].

We expect that if families were sampled with equal probability p, the family abundance in the metagenome $a^{MG}(f,m)$ and the total number of domains would be reduced by the same fraction. The selection of families with exponent $\beta_f \simeq 1$, not only estimates reliable the average genome size, but also confirms our prediction, $a^{MG}(f,m)/A_f$ indeed gives values that dispose around the total number of domains in the metagenome (referred to as the 'metagenome size') in every project (Fig.4.13).

Unlike for simulated metagenomes, the genome size distribution $P_m(n)$ is not known and we are forced to make assumption on its shape. The simplest hypothesis is using the exponential of eq. 4.16 for the transformed size distribution $G_m(\beta)$. Also assuming that family abundances in a metagenome are equally sampled with a constant probability p, the rescaled abundance is on

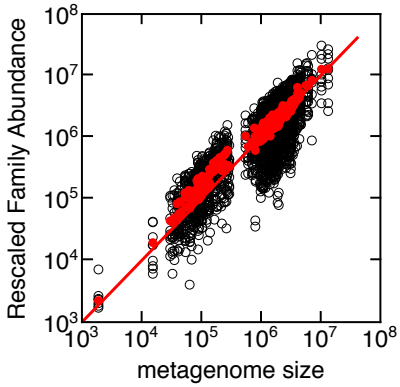

Figure 4.13: Each selected family with exponent $\beta \sim 1$ has a rescaled abundance that reproduces the size of sampled metagenomes. The graph shows for each metagenome, the rescaled family abundance $a^{MG}(f,m)/A_f$ (y-axis) as a function of the metagenome size $\sum_{f\in m} a^{MG}(f,m)$ (x-axis). The black dots correspond to the rescaled family abundance calculated with the 9 selected families with exponent $\beta \sim 1$. Red points represent the averaged value of $N\Gamma(f,m)$ over the 9 families and the red line is the bisecting line. The plot is in log-log scale in order to better visualize the small metagenomes. In this project, the families places around the bisecting line, which aligns well with their mean values. Data displayed in this plot are samples from the human gut microbioma collected in five different projects: 12 three-months old infants (ERP001038 [74]) , 147 samples from 145 70-years old european women with normal, impaired or diabetic glucose control (ERP002469 [39]), 18 samples from 6 European families, each one composed by two twins and the mother 9 of which are obese, 6 lean and 3 overweight (SRP000319 [80]), 45 samples from patients with diarrhea during the 2011 outbreak of Shiga-toxigenic Escherichia coli (STEC) O104:H4 in Germany (ERP001956 [50]), 19 samples, twelve of them are from patients with Chron's disease (SRP002423 [66]).

average

$$\langle N_{tot}\Gamma(f,m)\rangle = \frac{pa^{MG}(f,m)}{A_f} \sim pN_{tot}G_m(\beta_f) \tag{4.20}$$

implying that, in lin-log scale, $\langle N_{tot}\Gamma(f,m)\rangle$ describes the same curve of $N_{tot}G_m(\beta_f)$, only shifted vertically by a constant value $\log p$. Since the assumed exponential nature of $G_m(\beta)$, we expect that the logarithms of the averages of $N_{tot}\Gamma(f,m)$ binned by exponent β describe a straight line in β

$$\log(\langle N_{tot}\Gamma(f,m)\rangle) \sim \log(N_s) + \beta_f \log\langle n\rangle \tag{4.21}$$

with the logarithm of the average genome size $\langle n\rangle$ as slope and the logarithm of the sampled number of genomes $N_s = pN_{tot}$ as intercept. Figure 4.11 confirms this prediction. From each empirical metagenome among the 248 considered, we have obtained these two parameters by fitting the straight line, following the same procedure described in section 4.3.2 and then averaging the values N_s and $\langle n\rangle$ obtained for different occurrence cutoffs. The assumption made on the sampling of families implies that also the total size of the metagenome should be reduced by a fraction p. Therefore, applying the definition of mean, the average sampled metagenome size n_s is written as

$$\langle n_s\rangle = pn_m = pN_{tot}\langle n\rangle \tag{4.22}$$

This necessary condition is verified observing the product of the values N_s and $\langle n \rangle$ obtained with the fit reproduces well the size of the metagenome (Fig.4.14). This fact is non-trivial, since the

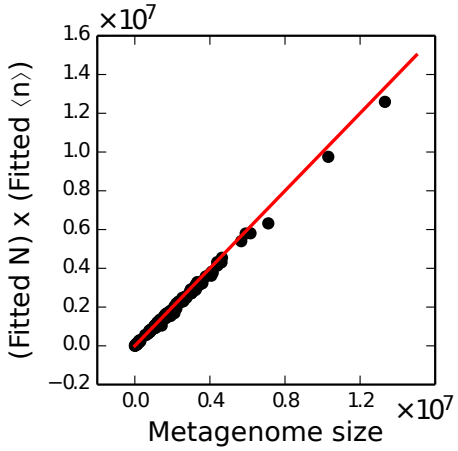

Figure 4.14: The number of genomes N and the average genome size $\langle n \rangle$ obtained by the fit reproduce well the metagenome size. The product of average genome size $\langle n \rangle$ and number of genomes N obtained by the fit (y-axis) is plotted against the metagenome size (x-axis) for each metagenome (black dots), aligning well with the line x=y (red line). This plot displays all metagenomes from all the examined EBI projects [39, 50, 66, 74, 80].

fitted parameters depend only on the selected scaling families (437 families over the 8675 in reference genomes) and do not depend directly on the total number of domains in the metagenome. Moreover, the accordance improved in every project by reducing the range of occurrence cutoffs to $[0.75, 0.79]$, which has been thus adopted to calculate N_s and $\langle n \rangle$.

Figure 4.14 is an important indication that the average genome size and the number of genomes in real metagenomes are estimated reliably. To further test the validity of our method, we focused on five metagenomics projects (details about how we retrieved empirical data are described in section 4.2).

Each project analyzes a set of human fecal samples characterized by different metabolic diseases (type 2 diabete in project ERP002469 [39], diarrhea caused by Shiga-toxigenic Escherichia coli (STEC) O104:H4 in project ERP001956 [50], Crohn's disease in project SRP002423 [66], obesity in project SRP000319 [80]) or by different diet (infants fed with breast milk or formula in project ERP001038 [74]). Examining the taxonomic annotations provided by the EBI database, we observed that the taxonomic composition of samples shows that projects whose samples contain a higher percentage of Firmicutes, exhibit a lower average genome size (Fig. 4.15) and this is consistent with reports that Firmicutes possess smaller genomes than Bacteroidetes [54]. Specifically, the average genome sizes show an overall, roughly linear decrease with the relative abundance of Firmicutes in metagenomes. Since a high number of 16S sequences are not assigned to a specific phylum (an average of 48% in SRP002423 samples), by calculating the relative abundances we are assuming that the proportions are kept for assigned sequences. The diversity of the sampling between projects and the clear diversity of taxonomic composition observed may depend partially

on the different DNA extraction protocols used by each study, as pointed out in [59]. The next paragraph will therefore analyze the diversity of metagenomic samples within the projects, where the protocols applied are the same.

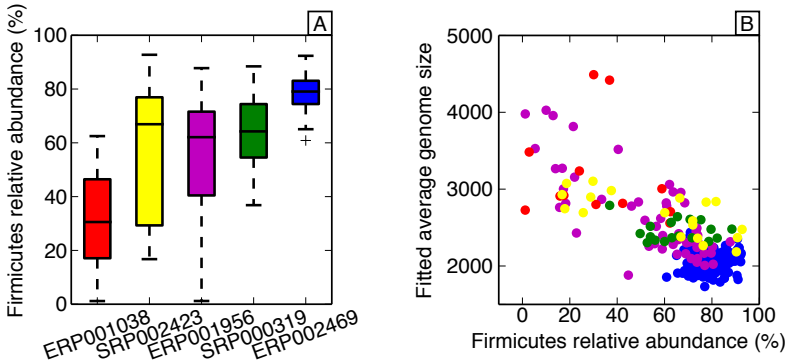

Figure 4.15: The estimated average genome size reflects the taxonomic composition of metagenomes across projects. (A) Boxplot of the estimated relative abundance of Firmicutes in each of the five projects analyzed, here indicated by their EBI accession code. (B) Roughly linear correlation between the average genome size (y-axis) and the relative abundance of Firmicutes (x-axis). Metagenomes of the same project have the same color: ERP001038 [74] in red, SRP002423 [66] in yellow, ERP001956 [50], in purple, SRP000319 [80] in green and ERP002469 [39] in blue.

As already mentioned in the introduction of the Chapter, the average genome size is an useful observable to compare different metagenomes. We selected three out of the five project to check if the average genome size could distinguish metagenome samples from individuals with different health conditions or diet. We didn't analyze project SRP002423 [66] because it is poorly documented and project ERP001956 [50] because it employs two different protocols for DNA extraction. Before discussing the result, the following paragraphs presents the details of the analyzed projects.

Project ERP002469 [39] lists 147 metagenomes relative to the gut microbiome of 145 70-year-old european women with normal, impaired or diabetic glucose control (2 samples have been analyzed twice by the EBI pipeline). Among the metagenomes, 50 belong to patients with impaired glucose control (IGT), 43 with normal glucose control (NGT) and 53 with type 2 diabetes (T2D). The use of metagenomics analysis allows to develop a mathematical model that predicts which women with impaired glucose tolerance have a diabetes-like metabolism. Diabetes, as well as other metabolic diseases, is influenced by socio-demographic and environmental factors more than by human genetics, for this reason the analysis of the gut microbiota as an environmental factor reveals to be particularly successful.

The 18 samples of project SRP000319 [80] derive from 6 European families, each one composed by two twins and the mother. The patients are divided in three groups: 9 obese, 6 lean and 3 overweight. This project is focused on determining how host genotype, environmental exposure and host adiposity influence the gut microbiome. The emerging result is that sampled individuals share a "core microbiome" at the gene level and deviation from this core determine different

physiological states.

The last project examined is ERP001038. There are only 12 metagenomes, which are relative to the fecal samples of 3 month-old infants differently fed: 6 of them were breast fed (BF) while the other 6 were exclusively nourished with a formula (FF) [74]. The study provides evidence that differences in diet can affect, via gut colonization, host expression of genes associated with the innate immune system. This study, published in April 2012, does not contradict a more recent study about the taxonomic composition of infant gut microbiota [81]. As a matter of fact, the taxonomic composition of infant fecal samples shows a high level of Actinobacteria.

For what concerns samples from project SRP000319, the average genome sizes assume similar values between obese and lean patients, exhibiting instead particularly high values in overweight ones (Fig.4.16A). The low statistic, however, does not allow to infer a clear trend. Similarly, in the case of project ERP002469, the average genome size has close values between patients with variable glucose control, exhibiting comparable distributions (Fig.4.16C). In addition, the average genome size does not show a clear correlation with the body mass index (BMI), spanning different values evenly both in case of lean patients (18.5 < BMI < 24.9) and obese ones (BMI > 30). Figure 4.16B shows that the average genome size distribution allows to distinguish between breast-feed infants (BF) and formula-feed infants (FF). Specifically, the metagenomes of BF infants can have extremely variable average genome sizes while the ones FF have a more peaked distribution (Fig.4.16B). This result is in agreement with the observation made by the original study [74], which showed that BF samples had a more heterogeneous phylogenetic composition than FF, a larger fraction of Bacteroidetes and a smaller fraction of Firmicutes.

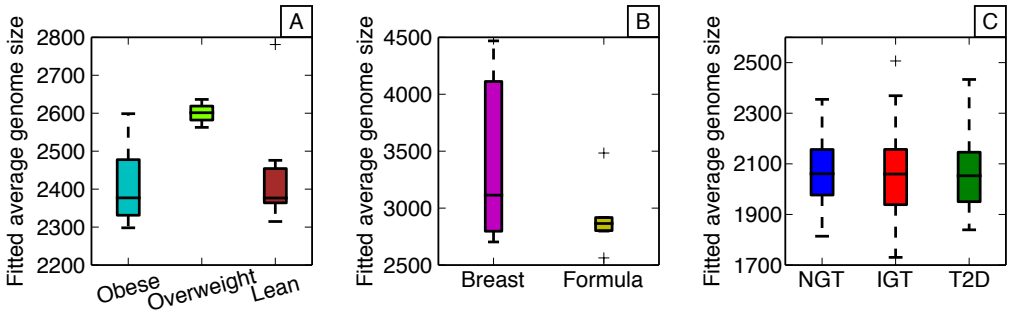

Figure 4.16: The average genome size in the gut microbiome depends strongly on the diet, diabetes and obesity on the contrary do not influence the average genome size. (A) Distributions of the average genome sizes of 18 human gut microbiomes deriving from project SRP000319 and divided in 9 obese patients (blue box), 3 overweight (green box) and 6 lean (brown box). (B) Distributions of the average genome sizes of 12 stool metagenomes deriving from project ERP001038 and divided by diet: 6 breast fed samples (magenta box) and 6 formula fed samples (yellow box). The metagenomes of breast-fed infants show a highly variable average genome size while the ones fed with a formula have a more specific value. (C) Distributions of the average genome sizes of 147 stool metagenomes deriving from project ERP002469 and divided by glucose control condition: 43 with normal glucose control (NGT, blue box), 50 with impaired glucose control (IGT, red box) and 53 with Type 2 diabetes (T2D, green box).

Only in the case of project ERP002469 [39], we found an independent estimation of the average

genome size (AGS) of samples carried out by Nayfach [59]. Figure 4.17 shows the scatter plot between our estimation of the average genome size and the one from [59]. The two sets of values are linearly correlated, with a Pearson correlation coefficient equal to 0.65. However, the AGS calculated fitting equation 4.21 is systematically lower than the one from Nayfach. The fact that Nayfach's AGS has been converted from basepairs to number of domains may have influenced this result.

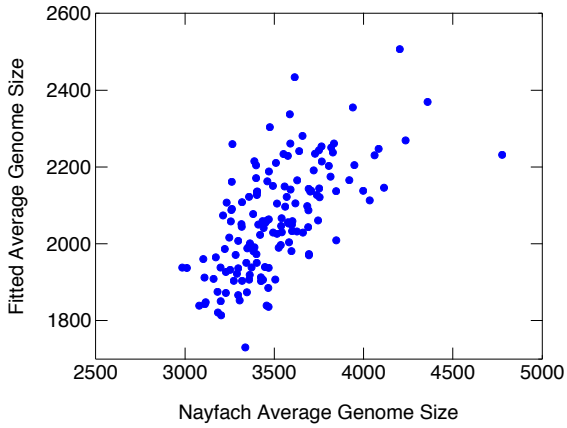

Figure 4.17: Comparison between the average genome size calculated fitting equation 4.21 and the average genome size derived in [59] for metagenomes of project ERP002469. The data displayed are metagenomic samples from project ERP002469, that is 147 fecal samples from women with normal, impaired or diabetic glucose control. Values of the average genome size calculated fitting equation 4.21 are linearly correlated with estimates of the average genome size from [59]. Both sets of values are measured in number of domains.

4.6 Conclusions

The abundance of different domain families in genomes is constrained by family-specific scaling laws. This result (fully exploited in Chapter 2) affects the abundance of domain families in a metagenome leading to the definition of a new observable which acts as a metagenome signature. The functional form of the newly defined metagenome invariant reflects the composition of the sample and when evaluated at integer scaling exponents, it gives access to the moments of the size distribution of genomes. In case of uniform or peaked genome size distribution, analytical calculations are possible and we can predict the theoretical value of the total number of genome in the sample, the mean and the variance of the genome size distribution.

We tested our results for simulated metagenomes, produced by random linear combinations of a set of $O(1000)$ reference genomes. The analysis supports the accuracy of theoretical predictions, but also highlighted their limitations. The estimation of the original number of microbes mixed in the sample and the average genome size are correct, however the same do not hold for the variance of the genome sizes. Only under certain conditions it is possible to access the variance, that is when

the range of scaling exponent are restricted to values equal to 2. Unfortunately high fluctuations of the rescaled abundance for large exponents make the estimate of the variance unreliable in most cases.

Finally we employed the tools developed in the first sections on 248 human gut microbiomes. Our theoretical predictions confirmed their validity, indeed the metagenome invariant for empirical samples shows the expected exponential dependency on the scaling exponent. Analyzed data show the evidence of domains sampling, that results in changes of the original abundances of families. Although this phenomenon does not follow the expected Poisson distribution, the reduction of domain abundances is uniform among families, thus limiting the change in the functional form of the scaling to a prefactor. This fact preserves the role of the rescaled abundance as signature of the composition of the metagenome, allowing in particular to obtain an estimate of its original average genome size. The comparison of these estimates evidenced the diversity between samples, reflecting the differences in the effective relative abundance of a particular class of bacteria.

Chapter 5

Conclusions and perspectives

In conclusion, the common underlying trait of the investigations described in this thesis is the representation of genomes as component systems. i.e. systems where modules, in our case the protein domains, can occur in different realizations, the genomes, with varying abundance. Such representation is useful to highlight several invariants found in the structure of the protein-coding part of genomes [16, 30, 31], but may also be useful for other [8, 22, 51, 65]. Example of other systems that can successfully be represented as component systems are ecosystems (where species are components), texts, software architectures (e.g. programs or operating systems), and in general all projects involving clear modules such as houses, lego sets, IKEA boxes, etc. Thus, the quantitative invariants and the theoretical tools developed here may be useful beyond genomes. Indeed, several notable quantitative laws can be identified in the composition of component systems of very different nature. For example, in linguistics, the notorious "Zipf's law" [87] describing the word frequency distribution (or its equivalent rank plot) in a text has been the subject of extensive investigations [67]. The existence of quantitative "universal" laws in texts may in principle provide insights on the cognitive mechanisms of text production, and can have practical applications in data mining and data search techniques [1].

While unifying traits may be common to different component systems, most investigations will be interested in the specificities leading to the systems peculiar architecture and behavior. Thus, we need to have a clear idea of the general behavior of component systems in general cases. This is by itself a challenging task, as such systems show a large degree of non-trivial universal properties [1, 42, 16] that could in principle affect the occurrence statistics. For example, the heterogeneous usage of different components, can be seen as a hallmark of the complexity of a component However, the ubiquity of this emergent behavior raises the question of whether (and to what extent) empirical laws like Zipf's law are pervasive statistical patterns that transcend system-specific mechanisms [42, 5]. In this spirit, the analysis of radically different systems can help the

discovery of patterns that descend from pure statistical effects or general principles [5, 65].

The systematic analysis of the abundance of domain families across a large sample of bacterial genomes performed in Chapter 2, revealed that family abundances increase as a power-law of the genome size. It is unclear whether other component systems may also exhibit this peculiar statistical laws. Surely, this kind of behavior is not reported for texts, which have been studied extensively, and therefore is likely not to hold in that context. With the help of a null model, we proved that these scaling laws are not simply due to sampling effects, thus bringing evidence against the hypothesis of combinatorially neutral scaling. The existence of family-specific scaling laws opens new perspectives on the evolutionary constraints that regulate the composition of genomes, and in particular on the interplay between domain families and their functions. In addition, the observed heterogeneity in the values of the scaling exponent across families in the same functional category, may provide new methods to suggest the need to revise or refine the functional annotations of protein domains.

Component systems, regardless of the field to which they belong, are characterized by a series of quantitative laws. A very ambitious goal is to unify all the observed regularities under a robust theoretical framework able to explain the mechanisms causing their emergence. In line with this idea, Chapter 3 presents a positive model that shows how the existence of a dependency structure linking components is responsible for the zipfian rank-frequency relation and also for the sublinear scaling of the number of unique components (Heaps' law). Mean-field calculations recovered the detailed structure of Heaps' law, i.e. the existence of three distinct scaling regimes and suggested the stretched exponential function as a good approximation of Heaps' law. Simulations based on our model matched the analytical predictions. Our analysis confirms the central role of dependency structures in shaping the properties of component systems. To be able to reach a deep understanding of empirical systems, the network should include constraints specific to the field examined, which are usually very hard to encode.

Finally, we studied (Chapter 4) the family abundance profiles in a metagenome, i.e. the sets of sequences found in a microbial ecosystem. In this study, we defined a different component system, i.e. the set of all the families found in the same ecosystem, with their abundances. However, since the defining modules (components) are identical to the one used for genomes, we could exploit the knowledge of the invariants valid in comparative genomics to study this system. In particular, we asked how the family-specific scaling laws valid for genomes translate into invariant quantities for the microbial community. The key result is that the rescaled abundance of a domain family in a metagenome has a functional form that is determined by the genome size composition of the metagenome.

Using both simulated and real metagenomes, we were able to calculate the total number of genomes that are combined in a metagenomic sample and more interestingly the average genome size. The average genome size is not readily available in metagenome studies, since all the DNA is pooled together and fragmented. Clearly, reconstructing this information is useful for comparing metagenomes sampled in different environments or within the same one. The newly discovered metagenomic invariant theoretically could give access to the moments of the genome size distribution. The fact that scaling exponents range up to 2, makes it impossible to estimate moments higher

than the second. We also encountered major difficulties even with the second moments, preventing us to evaluate the variance of the genome sizes. This is likely due to deviations of the rescaled family abundance, which are of the same magnitude in $\langle n^2 \rangle$. One possible way to reduce the fluctuations could be to disentangle the interdependency between the exponent and the prefactor in the family scaling laws.

Appendices

Appendix A

Supplementary tables

Table A.1: Scaling exponent of functional categories. The table reports the scaling exponent β_c of all functional categories examined, both for superfamilies (SUPFAM column) and clans (PFAM column). The error associated with the exponent is calculated as the root mean square deviation of the logarithm of the category abundance across all genomes from the estimated scaling law.

cat. code	category name	$A_c \pm \sigma_{A_c}$ (Supfam)	$\beta_c \pm \sigma_{\beta_c}$ (Supfam)	$A_c \pm \sigma_{A_c}$ (Pfam)	$\beta_c \pm \sigma_{\beta_c}$ (Pfam)
A	RNA binding, met./tr.	0.8 ± 0.5	0.4 ± 0.1	1.7 ± 0.2	0.21 ± 0.10
B	Chromatin structure	0.019 ± 0.099	0.6 ± 0.3	$--$	$--$
C	Energy	0.08 ± 0.03	0.8 ± 0.1	0.02 ± 0.01	0.94 ± 0.09
CA	E-transfer	0.00002 ± 0.01565	1.8 ± 0.3	0.00001 ± 0.02549	1.73 ± 0.34
CB	Photosynthesis	0.002 ± 0.051	0.9 ± 0.4	0.02 ± 0.08	0.56 ± 0.28
D	Cell cycle, Apoptosis	0.0007 ± 0.0272	1.2 ± 0.2	$--$	$--$
E	Amino acids m/tr	0.05 ± 0.12	0.9 ± 0.2	0.0021 ± 0.052	1.09 ± 0.17
EA	Nitrogen m/tr	0.000004 ± 0.009263	1.8 ± 0.2	$--$	$--$
F	Nucleotide m/tr	1.4 ± 0.3	0.5 ± 0.1	0.15 ± 0.08	0.70 ± 0.17
G	Carbohydrate m/tr	0.02 ± 0.05	1.0 ± 0.2	0.0003 ± 0.0372	1.38 ± 0.31
GA	Polysaccharide m/tr	0.005 ± 0.039	1.1 ± 0.2	0.003 ± 0.022	1.13 ± 0.20
H	Coenzyme m/tr	0.12 ± 0.05	0.9 ± 0.1	0.0009 ± 0.0204	1.27 ± 0.18
HA	Small molecule binding	0.21 ± 0.04	0.9 ± 0.1	0.55 ± 0.09	0.74 ± 0.06
HD	Receptor activity	0.0002 ± 0.0425	1.3 ± 0.5	0.3 ± 0.1	0.25 ± 0.20

HE	Ligand binding	0.3 ± 0.1	0.3 ± 0.1	0.0007 ± 0.0301	1.19 ± 0.31
I	Lipid m/tr	0.0009 ± 0.0176	1.2 ± 0.2	0.0002 ± 0.0135	1.31 ± 0.20
IA	Phospholipid m/tr	0.02 ± 0.04	0.6 ± 0.3	0.004 ± 0.027	0.77 ± 0.25
J	Translation	39.131 ± 0.006	0.16 ± 0.03	9.71 ± 0.02	0.17 ± 0.05
K	Transcription	0.02 ± 0.03	0.9 ± 0.1	0.003 ± 0.033	0.98 ± 0.19
L	DNA replication/repair	2.01 ± 0.05	0.5 ± 0.1	0.46 ± 0.04	0.63 ± 0.07
LA	DNA-binding	0.0007 ± 0.0143	1.5 ± 0.1	0.0009 ± 0.016	1.49 ± 0.13
LB	RNA processing	0.9 ± 1.3	0.2 ± 0.1	4.89 ± 0.08	0.07 ± 0.12
M	Cell envelope m/tr	0.02 ± 0.06	0.7 ± 0.4	−−	−−
MA	Cell adhesion	0.0006 ± 0.0310	1.3 ± 0.3	0.0001 ± 0.0354	1.49 ± 0.29
N	Cell motility	0.02 ± 0.07	0.7 ± 0.3	0.10 ± 0.08	0.48 ± 0.20
O	Protein modification	0.01 ± 0.03	1.1 ± 0.1	0.006 ± 0.025	0.99 ± 0.13
OA	Proteases	0.04 ± 0.02	1.0 ± 0.1	0.02 ± 0.02	1.01 ± 0.10
OB	Kinases/phosphatases	0.0008 ± 0.0321	1.3 ± 0.2	0.0007 ± 0.0479	1.02 ± 0.31
P	Ion m/tr	0.005 ± 0.023	1.3 ± 0.1	0.003 ± 0.034	1.04 ± 0.17
Q	Secondary metabolism	0.00009 ± 0.01176	1.5 ± 0.2	0 ± 0.02	2.01 ± 0.29
R	General	0.001 ± 0.013	1.2 ± 0.2	0.00002 ± 0.0164	1.61 ± 0.23
RA	Redox	0.008 ± 0.026	1.2 ± 0.1	0.006 ± 0.036	1.19 ± 0.13
RB	Transferases	0.04 ± 0.02	1.1 ± 0.1	0.01 ± 0.02	1.13 ± 0.09
RC	Other enzymes	0.07 ± 0.03	1.1 ± 0.1	0.04 ± 0.02	1.13 ± 0.06
RD	Protein interaction	0.02 ± 0.07	1.0 ± 0.2	0.04 ± 0.05	0.93 ± 0.19
RF	Transport	0.02 ± 0.04	1.1 ± 0.2	0.004 ± 0.023	1.10 ± 0.16
S	Unknown function	0.04 ± 0.02	1.0 ± 0.1	0.03 ± 0.04	0.89 ± 0.13
SB	Toxins/defence	0.001 ± 0.068	1.1 ± 0.3	0.00006 ± 0.04413	1.28 ± 0.29
T	Signal transduction	0.0002 ± 0.0303	1.6 ± 0.2	0.00007 ± 0.03247	1.74 ± 0.20
TA	Other regulatory function	0.005 ± 0.035	1.1 ± 0.2	0.02 ± 0.05	0.55 ± 0.24

Table A.2: Scaling exponent of superfamilies from the SUPERFAMILY database. The abundance of a (super)family scales as a power law of the genome size with family-dependent scaling exponents β_i. Each row corresponds to a domain family and shows its scaling exponent along with its error and the category to which the family belongs (category code). Families corresponding to the same functional category are ordered in decreasing order of abundance. The error associated with the exponent is calculated as the root mean square deviation of the logarithm of the category abundance across all genomes from the estimated scaling law.

cat. code	family name	$A_i \pm \sigma_{A_i}$	$\beta_i \pm \sigma_{\beta_i}$
A	Alpha-L RNA-binding motif	1.07 ± 1.71	0.2 ± 0.1
A	PUA domain-like	0.01 ± 0.04	0.7 ± 0.2
C	6-phosphogluconate dehydrogenase C-terminal domain-like	0.0005 ± 0.0186	1.2 ± 0.2
C	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	0.004 ± 0.037	1.0 ± 0.2
C	Phosphoenolpyruvate/pyruvate domain	0.0004 ± 0.0185	1.0 ± 0.2
C	SIS domain	0.004 ± 0.037	0.7 ± 0.2
C	LeuD/IlvD-like	0.002 ± 0.034	0.9 ± 0.2
C	Enolase C-terminal domain-like	0.02 ± 0.07	0.8 ± 0.2
C	Transmembrane di-heme cytochromes	0.003 ± 0.036	0.5 ± 0.3
C	Aconitase iron-sulfur domain	0.004 ± 0.046	0.5 ± 0.2
C	Cytochrome c oxidase subunit I-like	0.06 ± 0.09	0.5 ± 0.2
C	UDP-glucose/ GDP-mannose dehydrogenase C-terminal domain	0.03 ± 0.03	0.5 ± 0.2
C	Citrate synthase	0.009 ± 0.061	0.6 ± 0.2
C	PEP carboxykinase-like	0.24 ± 0.07	0.2 ± 0.2
C	Cytochrome c oxidase subunit III-like	0.07 ± 0.07	0.4 ± 0.2
C	PK C-terminal domain-like	0.17 ± 0.06	0.2 ± 0.1
C	Enzyme I of the PEP:sugar phosphotransferase system HPr-binding (sub)domain	0.05 ± 0.05	0.4 ± 0.2
CA	Cytochrome c	0.002 ± 0.067	1.0 ± 0.5
CA	Acyl-CoA dehydrogenase C-terminal domain-like	0.0 ± 0.02	2.0 ± 0.4
CA	FMN-dependent nitroreductase-like	0.004 ± 0.040	0.8 ± 0.3
CA	ISP domain	0.0003 ± 0.0495	1.2 ± 0.3
CA	Sulfite reductase hemoprotein (SiRHP), domains 2 and 4	0.002 ± 0.048	0.9 ± 0.2
CA	Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain	0.01 ± 0.03	0.7 ± 0.2
CB	PRC-barrel domain	0.008 ± 0.069	0.7 ± 0.3
D	Rhodanese/Cell cycle control phosphatase	0.0006 ± 0.0242	1.1 ± 0.3
E	ACT-like	0.02 ± 0.17	0.8 ± 0.2

E	Tryptophan synthase beta subunit-like PLP-dependent enzymes	0.002 ± 0.038	1.0 ± 0.2
E	Carbamate kinase-like	0.08 ± 0.10	0.5 ± 0.1
E	PLP-binding barrel	0.02 ± 0.07	0.7 ± 0.1
E	Glutamine synthetase/guanido kinase	0.02 ± 0.04	0.7 ± 0.2
E	L-aspartase-like	0.01 ± 0.03	0.7 ± 0.2
E	Diaminopimelate epimerase-like	0.04 ± 0.09	0.5 ± 0.2
E	Alanine racemase C-terminal domain-like	0.03 ± 0.07	0.5 ± 0.2
E	Aspartate/glutamate racemase	0.3 ± 0.2	0.3 ± 0.2
E	Arginase/deacetylase	0.005 ± 0.035	0.8 ± 0.2
E	Aspartate/ornithine carbamoyltransferase	0.21 ± 0.07	0.3 ± 0.1
E	Serine metabolism enzymes domain	0.11 ± 0.04	0.3 ± 0.2
E	Chorismate mutase II	0.06 ± 0.05	0.4 ± 0.2
EA	RmlC-like cupins	0.000006 ± 0.016133	1.8 ± 0.2
F	Ribonuclease H-like	0.06 ± 0.05	0.7 ± 0.3
F	Adenine nucleotide alpha hydrolases-like	0.008 ± 0.030	0.9 ± 0.1
F	Nucleotidylyl transferase	0.2 ± 0.2	0.17 ± 0.05
F	PRTase-like	0.03 ± 0.08	0.4 ± 0.1
F	Nucleotidyltransferase	0.6 ± 0.3	0.7 ± 0.2
F	Pseudouridine synthase	0.1 ± 0.2	0.3 ± 0.1
F	Ribulose-phoshate binding barrel	0.11 ± 0.05	0.5 ± 0.2
F	Tetrahydrobiopterin biosynthesis enzymes-like	0.17 ± 0.07	0.4 ± 0.2
F	Purine and uridine phosphorylases	0.5 ± 0.3	0.3 ± 0.2
F	Nucleotidyltransferase substrate binding subunit/domain	0.06 ± 0.04	0.2 ± 0.2
F	Nicotinate/Quinolinate PRTase C-terminal domain-like	0.08 ± 0.10	0.4 ± 0.2
F	Nucleoside phosphorylase/ phosphoribosyltransferase catalytic domain	0.09 ± 0.06	0.3 ± 0.2
F	Nucleoside phosphorylase/ phosphoribosyltransferase N-terminal domain	0.04 ± 0.04	0.4 ± 0.2
G	(Trans)glycosidases	0.0002 ± 0.0441	1.4 ± 0.4
G	Aldolase	0.007 ± 0.043	0.9 ± 0.2
G	Phosphoglucomutase, first 3 domains	0.3 ± 0.1	0.4 ± 0.1
G	Galactose-binding domain-like	0.007 ± 0.103	0.8 ± 0.5
G	Six-hairpin glycosidases	0.0003 ± 0.0413	1.2 ± 0.4
G	Duplicated hybrid motif	0.08 ± 0.05	0.5 ± 0.2
G	Xylose isomerase-like	0.001 ± 0.062	1.0 ± 0.3
G	Carbohydrate phosphatase	0.02 ± 0.05	0.6 ± 0.2
G	HIT-like	0.01 ± 0.04	0.6 ± 0.2
G	Phosphoglucomutase, C-terminal domain	0.11 ± 0.07	0.4 ± 0.1

G	PK beta-barrel domain-like	0.001 ± 0.019	0.9 ± 0.2
G	HPr-like	0.2 ± 0.1	0.2 ± 0.2
GA	UDP-Glycosyltransferase/glycogen phosphorylase	0.007 ± 0.037	1.0 ± 0.2
GA	Pectin lyase-like	0.0001 ± 0.0516	1.3 ± 0.4
GA	Glycosyl hydrolase domain	0.005 ± 0.053	0.8 ± 0.3
GA	Barwin-like endoglucanases	0.02 ± 0.03	0.5 ± 0.2
H	Glutathione synthetase ATP-binding domain-like	0.01 ± 0.03	0.9 ± 0.1
H	Acyl-CoA dehydrogenase NM domain-like	0.00 ± 0.02	2.0 ± 0.4
H	PreATP-grasp domain	0.04 ± 0.06	0.7 ± 0.1
H	Single hybrid motif	0.01 ± 0.03	0.8 ± 0.2
H	FMN-binding split barrel	0.0003 ± 0.0429	1.2 ± 0.2
H	Riboflavin synthase domain-like	0.001 ± 0.025	1.0 ± 0.2
H	Succinyl-CoA synthetase domains	0.07 ± 0.06	0.5 ± 0.2
H	YrdC/RibB	0.04 ± 0.03	0.5 ± 0.2
H	Molybdenum cofactor biosynthesis proteins	0.03 ± 0.06	0.6 ± 0.2
H	Dihydrofolate reductase-like	0.02 ± 0.07	0.6 ± 0.2
H	UROD/MetE-like	0.05 ± 0.08	0.5 ± 0.3
H	Dihydropteroate synthetase-like	0.07 ± 0.08	0.4 ± 0.2
H	Cobalamin (vitamin B12)-binding domain	0.06 ± 0.06	0.4 ± 0.3
H	Activating enzymes of the ubiquitin-like proteins	0.07 ± 0.04	0.4 ± 0.2
H	Nicotinate/Quinolinate PRTase N-terminal domain-like	0.01 ± 0.05	0.4 ± 0.2
H	Glutamine synthetase, N-terminal domain	0.3 ± 0.1	0.6 ± 0.2
H	Peptide deformylase	0.06 ± 0.12	0.2 ± 0.2
H	RibA-like	0.2 ± 0.1	0.4 ± 0.2
H	MoeA C-terminal domain-like	0.08 ± 0.06	0.2 ± 0.2
H	ApbE-like	0.001 ± 0.017	0.3 ± 0.2
HA	P-loop containing nucleoside triphosphate hydrolases	0.001 ± 0.010	0.71 ± 0.08
HA	NAD(P)-binding Rossmann-fold domains	0.01 ± 0.03	1.4 ± 0.1
HA	FAD/NAD(P)-binding domain	0.002 ± 0.042	1.3 ± 0.2
HA	Thiamin diphosphate-binding fold (THDP-binding)	0.005 ± 0.026	0.9 ± 0.1
HA	FAD-binding domain	0.02 ± 0.12	1.0 ± 0.2
HA	Nucleotide-binding domain	0.0002 ± 0.0441	0.8 ± 0.2
HA	Sensory domain-like	0.007 ± 0.043	0.7 ± 0.4
HD	Methyl-accepting chemotaxis protein (MCP) signaling domain	0.002 ± 0.080	1.0 ± 0.5
HD	PhoU-like	0.2 ± 0.1	0.3 ± 0.2
HE	TGS-like	0.3 ± 0.1	0.3 ± 0.1
I	Thioesterase/thiol ester dehydrase-isomerase	0.00006 ± 0.02319	1.5 ± 0.2
I	Probable ACP-binding domain of	0.0005 ± 0.0721	1.0 ± 0.3

	malonyl-CoA ACP transacylase		
I	Creatinase/prolidase N-terminal domain	0.03 ± 0.03	0.5 ± 0.2
I	Prokaryotic lipoproteins and lipoprotein localization factors	0.09 ± 0.06	0.4 ± 0.2
IA	PLC-like phosphodiesterases	0.02 ± 0.06	0.6 ± 0.2
J	Ribosomal protein S5 domain 2-like	1.8 ± 0.2	0.30 ± 0.07
J	Translation proteins	1.43 ± 0.09	0.26 ± 0.06
J	EF-G C-terminal domain-like	0.7 ± 0.2	0.30 ± 0.09
J	Sm-like ribonucleoproteins	0.005 ± 0.040	0.8 ± 0.3
J	Triger factor/SurA peptide-binding domain-like	0.3 ± 0.1	0.3 ± 0.2
J	Release factor	0.14 ± 0.12	0.2 ± 0.1
J	L30e-like	0.04 ± 0.03	0.3 ± 0.2
J	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain	0.2 ± 0.1	0.5 ± 0.2
J	S13-like H2TH domain	0.14 ± 0.06	0.3 ± 0.2
J	NusB-like	0.06 ± 0.04	0.3 ± 0.1
J	ClpS-like	0.08 ± 0.06	0.4 ± 0.1
J	Ribosome binding protein Y (YfiA homologue)	0.01 ± 0.01	0.3 ± 0.1
K	Tetracyclin repressor-like, C-terminal domain	0.03 ± 0.05	2.4 ± 0.3
K	LexA/Signal peptidase	0.3 ± 0.1	0.6 ± 0.2
K	Poly A polymerase C-terminal region-like	0.25 ± 0.10	0.2 ± 0.2
K	GreA transcript cleavage protein, N-terminal domain	0.10 ± 0.09	0.2 ± 0.1
K	CYTH-like phosphatases	2.87 ± 0.05	0.3 ± 0.1
L	Nucleic acid-binding proteins	0.0006 ± 0.0184	0.31 ± 0.07
L	DNA breaking-rejoining enzymes	0.2 ± 0.1	1.0 ± 0.3
L	Nudix	0.01 ± 0.06	1.2 ± 0.2
L	RuvA domain 2-like	0.004 ± 0.026	0.4 ± 0.1
L	Restriction endonuclease-like	0.01 ± 0.05	0.8 ± 0.3
L	DNA/RNA polymerases	0.005 ± 0.036	0.8 ± 0.2
L	DNA-glycosylase	0.02 ± 0.06	0.7 ± 0.2
L	DNase I-like	1.8 ± 0.2	0.8 ± 0.2
L	DNA polymerase III clamp loader subunits, C-terminal domain	0.5 ± 0.2	0.2 ± 0.1
L	Resolvase-like	0.07 ± 0.12	0.4 ± 0.4
L	Uracil-DNA glycosylase-like	0.02 ± 0.04	0.6 ± 0.2
L	GIY-YIG endonuclease	0.08 ± 0.05	0.4 ± 0.2
L	DNA ligase/mRNA capping enzyme, catalytic domain	0.007 ± 0.072	0.7 ± 0.2
L	HRDC-like	0.08 ± 0.07	0.4 ± 0.2
L	N-terminal domain of MutM-like DNA repair proteins	0.07 ± 0.07	0.4 ± 0.2
L	TRCF domain-like	0.00002 ± 0.03128	0.02 ± 0.03

LA	Winged helix DNA-binding domain	0.000001 ± 0.026499	1.8 ± 0.2
LA	Homeodomain-like	0.0002 ± 0.0262	2.2 ± 0.3
LA	lambda repressor-like DNA-binding domains	0.07 ± 0.12	1.4 ± 0.3
LA	C-terminal effector domain of the bipartite response regulators	0.000006 ± 0.010207	1.8 ± 0.2
LA	Periplasmic binding protein-like I	0.0001 ± 0.0287	1.4 ± 0.4
LA	Fatty acid responsive transcription factor FadR, C-terminal domain	0.000001 ± 0.028036	1.9 ± 0.3
LA	Glucocorticoid receptor-like (DNA-binding domain)	0.23 ± 0.09	0.4 ± 0.2
LA	TrpR-like	0.10 ± 0.06	0.4 ± 0.3
LA	Ribbon-helix-helix	0.006 ± 0.055	0.8 ± 0.3
LA	IHF-like DNA-binding proteins	0.2 ± 0.2	0.3 ± 0.3
LA	ParB/Sulfiredoxin	0.02 ± 0.04	0.6 ± 0.3
LA	KorB DNA-binding domain-like	0.05 ± 0.06	0.4 ± 0.3
LB	EPT/RTPC-like	0.2 ± 0.1	0.3 ± 0.1
M	OmpA-like	0.003 ± 0.047	0.9 ± 0.4
MA	vWA-like	0.0002 ± 0.0255	1.2 ± 0.3
MA	Pili subunits	0.010 ± 0.105	0.8 ± 0.4
MA	PGBD-like	0.01 ± 0.07	0.7 ± 0.3
MA	Hedgehog/DD-peptidase	0.05 ± 0.05	0.5 ± 0.2
O	ATPase domain of HSP90 chaperone/ DNA topoisomerase II/ histidine kinase	0.0001 ± 0.0209	1.5 ± 0.2
O	GroES-like	0.00002 ± 0.02546	1.6 ± 0.3
O	FKBP-like	0.06 ± 0.06	0.6 ± 0.2
O	Chaperone J-domain	0.07 ± 0.07	0.5 ± 0.2
O	Cyclophilin-like	0.003 ± 0.021	0.9 ± 0.2
O	Double Clp-N motif	0.02 ± 0.04	0.6 ± 0.2
O	HSP20-like chaperones	0.02 ± 0.04	0.6 ± 0.2
O	GroEL equatorial domain-like	0.2 ± 0.1	0.2 ± 0.2
O	GroEL apical domain-like	0.2 ± 0.2	0.2 ± 0.2
O	Peptide methionine sulfoxide reductase	0.25 ± 0.07	0.2 ± 0.2
O	GroEL-intermediate domain like	0.2 ± 0.1	0.2 ± 0.1
OA	ClpP/crotonase	0.003 ± 0.047	1.0 ± 0.2
OA	Zn-dependent exopeptidases	0.003 ± 0.034	1.0 ± 0.2
OA	Metallo-dependent phosphatases	0.002 ± 0.029	1.0 ± 0.2
OA	Metalloproteases ("zincins"), catalytic domain	0.007 ± 0.029	0.8 ± 0.3
OA	LuxS/MPP-like metallohydrolase	0.1 ± 0.2	0.5 ± 0.3
OA	Cysteine proteinases	0.002 ± 0.039	1.0 ± 0.3
OA	Bacterial exopeptidase dimerisation domain	0.0006 ± 0.0236	1.1 ± 0.3

OA	Trypsin-like serine proteases	0.001 ± 0.039	1.0 ± 0.3
OA	Creatinase/aminopeptidase	0.05 ± 0.03	0.5 ± 0.1
OA	HSP40/DnaJ peptide-binding domain	0.35 ± 0.09	0.3 ± 0.2
OA	DPP6 N-terminal domain-like	0.01 ± 0.11	0.6 ± 0.4
OA	Subtilisin-like	0.003 ± 0.060	0.8 ± 0.3
OA	Rhomboid-like	0.04 ± 0.07	0.5 ± 0.2
OA	Macro domain-like	0.10 ± 0.09	0.3 ± 0.2
OA	Tricorn protease N-terminal domain	0.07 ± 0.11	0.4 ± 0.2
OB	Protein kinase-like (PK-like)	0.0006 ± 0.0577	1.2 ± 0.3
OB	PP2C-like	0.005 ± 0.068	0.8 ± 0.3
OB	Phosphohistidine domain	0.009 ± 0.026	0.7 ± 0.2
OB	Phosphotyrosine protein phosphatases I	0.02 ± 0.04	0.6 ± 0.2
OB	Acylphosphatase/BLUF domain-like	0.15 ± 0.08	0.3 ± 0.2
P	Periplasmic binding protein-like II	0.00007 ± 0.02273	1.6 ± 0.3
P	MFS general substrate transporter	0.0003 ± 0.0379	1.4 ± 0.3
P	Multidrug resistance efflux transporter EmrE	0.0003 ± 0.0310	1.3 ± 0.3
P	HlyD-like secretion proteins	0.00006 ± 0.03092	1.5 ± 0.4
P	Ferritin-like	0.002 ± 0.016	1.0 ± 0.2
P	Cupredoxins	0.0009 ± 0.0273	1.1 ± 0.3
P	Calcium ATPase, transduction domain A	0.05 ± 0.08	0.6 ± 0.2
P	Calcium ATPase, transmembrane domain M	0.05 ± 0.07	0.6 ± 0.2
P	TrkA C-terminal domain-like	0.03 ± 0.08	0.6 ± 0.3
P	HMA, heavy metal-associated domain	0.1 ± 0.2	0.4 ± 0.2
P	Band 7/SPFH domain	0.06 ± 0.15	0.5 ± 0.2
P	Fe-S cluster assembly (FSCA) domain-like	0.06 ± 0.06	0.4 ± 0.2
P	Voltage-gated potassium channels	0.02 ± 0.04	0.6 ± 0.2
P	Magnesium transport protein CorA, transmembrane region	0.03 ± 0.05	0.5 ± 0.2
P	CorA soluble domain-like	0.04 ± 0.05	0.5 ± 0.2
P	Clc chloride channel	0.1 ± 0.2	0.3 ± 0.2
Q	Dimeric alpha+beta barrel	0.00 ± 0.02	2.1 ± 0.3
Q	Clavaminate synthase-like	0.00001 ± 0.03341	1.6 ± 0.3
Q	Concanavalin A-like lectins/glucanases	0.007 ± 0.086	0.8 ± 0.4
Q	Terpenoid synthases	0.01 ± 0.04	0.7 ± 0.2
Q	Homo-oligomeric flavin-containing Cys decarboxylases, HFCD	0.08 ± 0.04	0.4 ± 0.2
R	Bet v1-like	0.00002 ± 0.03706	1.5 ± 0.4
R	Helical backbone metal receptor	0.004 ± 0.051	0.9 ± 0.3
R	ADC-like	0.001 ± 0.050	1.0 ± 0.3
R	ARM repeat	0.02 ± 0.07	0.6 ± 0.3

R	Peripheral subunit-binding domain of 2-oxo acid dehydrogenase complex	0.2 ± 0.1	0.3 ± 0.2
R	Pentein	0.04 ± 0.07	0.4 ± 0.2
R	JAB1/MPN domain	0.19 ± 0.06	0.2 ± 0.2
RA	Thioredoxin-like	0.002 ± 0.034	1.1 ± 0.2
RA	4Fe-4S ferredoxins	0.02 ± 0.08	0.8 ± 0.4
RA	Metallo-hydrolase/oxidoreductase	0.002 ± 0.022	1.1 ± 0.2
RA	Glyoxalase/Bleomycin resistance protein/ Dihydroxybiphenyl dioxygenase	0 ± 0.01	2.1 ± 0.3
RA	ALDH-like	0.00003 ± 0.01758	1.5 ± 0.2
RA	2Fe-2S ferredoxin-like	0.001 ± 0.064	1.0 ± 0.3
RA	Flavoproteins	0.005 ± 0.044	0.9 ± 0.3
RA	alpha-helical ferredoxin	0.004 ± 0.063	0.9 ± 0.3
RA	FAD-linked reductases, C-terminal domain	0.00007 ± 0.04992	1.4 ± 0.3
RA	Formate/glycerate dehydrogenase catalytic domain-like	0.001 ± 0.026	1.0 ± 0.2
RA	NAD(P)-linked oxidoreductase	0.0001 ± 0.0495	1.3 ± 0.3
RA	Isocitrate/Isopropylmalate dehydrogenase-like	0.03 ± 0.04	0.6 ± 0.2
RA	Aminoacid dehydrogenase-like, N-terminal domain	0.02 ± 0.02	0.7 ± 0.1
RA	FAD/NAD-linked reductases, dimerisation (C-terminal) domain	0.02 ± 0.05	0.7 ± 0.2
RA	Formate dehydrogenase/DMSO reductase, domains 1-3	0.0009 ± 0.0403	1.0 ± 0.3
RA	Ferredoxin reductase-like, C-terminal NADP-linked domain	0.0009 ± 0.0463	1.0 ± 0.3
RA	Dehydroquinate synthase-like	0.01 ± 0.05	0.7 ± 0.3
RA	Inosine monophosphate dehydrogenase (IMPDH)	0.03 ± 0.05	0.5 ± 0.2
RA	Acid phosphatase/Vanadium-dependent haloperoxidase	0.009 ± 0.040	0.7 ± 0.2
RA	FAD-linked oxidases, C-terminal domain	0.0007 ± 0.0303	1.0 ± 0.3
RA	Succinate dehydrogenase/ fumarate reductase flavoprotein C-terminal domain	0.03 ± 0.02	0.5 ± 0.2
RA	LDH C-terminal domain-like	0.2 ± 0.1	0.3 ± 0.2
RA	FAD-linked oxidoreductase	0.04 ± 0.03	0.6 ± 0.2
RB	S-adenosyl-L-methionine-dependent methyltransferases	0.002 ± 0.038	0.9 ± 0.1
RB	PLP-dependent transferases	0.00003 ± 0.03022	1.2 ± 0.1
RB	Acyl-CoA N-acyltransferases (Nat)	0.01 ± 0.06	1.7 ± 0.2
RB	Nucleotide-diphospho-sugar transferases	0.004 ± 0.032	0.9 ± 0.2
RB	Class I glutamine amidotransferase-like	0.001 ± 0.080	1.0 ± 0.1
RB	CoA-dependent acyltransferases	0.001 ± 0.024	1.0 ± 0.5
RB	NagB/RpiA/CoA transferase-like	0.02 ± 0.05	1.1 ± 0.2
RB	TK C-terminal domain-like	0.0006 ± 0.0540	0.7 ± 0.2
RB	FabD/lysophospholipase-like	0.002 ± 0.029	1.1 ± 0.3

RB	Tetrapyrrole methylase	0.009 ± 0.055	1.0 ± 0.3
RB	Glycerol-3-phosphate (1)-acyltransferase	0.03 ± 0.03	0.7 ± 0.2
RB	Formyltransferase	0.04 ± 0.03	0.6 ± 0.1
RB	D-aminoacid aminotransferase-like PLP-dependent enzymes	0.02 ± 0.05	0.5 ± 0.2
RB	4'-phosphopantetheinyl transferase	0.2 ± 0.1	0.5 ± 0.2
RB	Methylated DNA-protein cysteine methyltransferase, C-terminal domain	0.004 ± 0.027	0.8 ± 0.2
RB	Methylated DNA-protein cysteine methyltransferase domain	0.02 ± 0.05	0.6 ± 0.2
RC	alpha/beta-Hydrolases	0.00005 ± 0.02412	1.6 ± 0.3
RC	Actin-like ATPase domain	0.12 ± 0.04	0.7 ± 0.1
RC	HAD-like	0.01 ± 0.08	0.9 ± 0.2
RC	Thiolase-like	0.0004 ± 0.0460	1.3 ± 0.3
RC	Radical SAM enzymes	0.06 ± 0.15	0.7 ± 0.3
RC	Acetyl-CoA synthetase-like	0.000002 ± 0.019005	1.9 ± 0.3
RC	Metallo-dependent hydrolases	0.0003 ± 0.0250	1.3 ± 0.2
RC	HD-domain/PDEase-like	0.04 ± 0.11	0.7 ± 0.3
RC	beta-lactamase/transpeptidase-like	0.005 ± 0.022	0.9 ± 0.2
RC	Trimeric LpxA-like enzymes	0.04 ± 0.06	0.7 ± 0.2
RC	Lysozyme-like	0.002 ± 0.027	1.0 ± 0.2
RC	Composite domain of metallo-dependent hydrolases	0.00008 ± 0.01467	1.4 ± 0.2
RC	N-terminal nucleophile aminohydrolases (Ntn hydrolases)	0.001 ± 0.032	1.1 ± 0.2
RC	Ribokinase-like	0.005 ± 0.041	0.9 ± 0.2
RC	Alkaline phosphatase-like	0.002 ± 0.039	1.0 ± 0.3
RC	DHS-like NAD/FAD-binding domain	0.0002 ± 0.0226	1.2 ± 0.2
RC	Phospholipase D/nuclease	0.007 ± 0.059	0.8 ± 0.2
RC	Glycoside hydrolase/deacetylase	0.0004 ± 0.0145	1.1 ± 0.2
RC	Cytidine deaminase-like	0.04 ± 0.06	0.6 ± 0.1
RC	LysM domain	0.4 ± 0.7	0.3 ± 0.4
RC	SGNH hydrolase	0.003 ± 0.056	0.9 ± 0.3
RC	PurM N-terminal domain-like	0.7 ± 0.2	0.2 ± 0.1
RC	PurM C-terminal domain-like	0.6 ± 0.2	0.2 ± 0.1
RC	Phosphoglycerate mutase-like	0.003 ± 0.038	0.9 ± 0.3
RC	Galactose mutarotase-like	0.04 ± 0.07	0.6 ± 0.3
RC	Carbon-nitrogen hydrolase	0.005 ± 0.019	0.8 ± 0.2
RC	PHP domain-like	0.02 ± 0.06	0.6 ± 0.2
RC	Enolase N-terminal domain-like	0.002 ± 0.057	0.9 ± 0.3
RC	Quinoprotein alcohol dehydrogenase-like	0.002 ± 0.051	0.9 ± 0.3
RC	all-alpha NTP pyrophosphatases	0.06 ± 0.12	0.5 ± 0.2
RC	FAH	0.00006 ± 0.02999	1.3 ± 0.3

RC	PFL-like glycyl radical enzymes	0.2 ± 0.3	0.3 ± 0.3
RC	Amidase signature (AS) enzymes	0.003 ± 0.066	0.8 ± 0.3
RC	Isochorismatase-like hydrolases	0.0002 ± 0.0469	1.2 ± 0.3
RC	L,D-transpeptidase catalytic domain-like	0.002 ± 0.036	0.9 ± 0.3
RC	Chorismate lyase-like	0.005 ± 0.076	0.8 ± 0.3
RC	MoCo carrier protein-like	0.03 ± 0.03	0.5 ± 0.2
RC	NAD kinase	0.03 ± 0.04	0.5 ± 0.2
RC	ADC synthase	0.07 ± 0.05	0.4 ± 0.2
RC	Folate-binding domain	0.02 ± 0.06	0.6 ± 0.2
RC	AraD-like aldolase/epimerase	0.007 ± 0.044	0.7 ± 0.2
RC	FMT C-terminal domain-like	0.2 ± 0.1	0.3 ± 0.2
RC	IlvD/EDD N-terminal domain-like	0.007 ± 0.070	0.7 ± 0.2
RC	Chelatase	0.06 ± 0.04	0.4 ± 0.2
RC	Aminomethyltransferase beta-barrel domain	0.009 ± 0.031	0.6 ± 0.2
RC	2-isopropylmalate synthase LeuA, allosteric (dimerisation) domain	0.16 ± 0.09	0.3 ± 0.2
RC	CNF1/YfiH-like putative cysteine hydrolases	0.3 ± 0.1	0.2 ± 0.2
RC	Nqo1 middle domain-like	0.3 ± 0.1	0.2 ± 0.2
RC	beta-carbonic anhydrase, cab	0.007 ± 0.049	0.7 ± 0.2
RC	N-acetylmuramoyl-L-alanine amidase-like	0.2 ± 0.2	0.3 ± 0.2
RC	post-HMGL domain-like	0.05 ± 0.06	0.4 ± 0.2
RC	Nqo1C-terminal domain-like	0.2 ± 0.1	0.2 ± 0.2
RC	DmpA/ArgJ-like	0.04 ± 0.05	0.4 ± 0.2
RC	LigT-like	0.10 ± 0.06	0.3 ± 0.2
RD	TPR-like	0.001 ± 0.051	1.2 ± 0.4
RD	FMN-linked oxidoreductases	0.002 ± 0.019	1.0 ± 0.1
RD	Nqo1 FMN-binding domain-like	0.13 ± 0.07	0.3 ± 0.2
RF	Multidrug efflux transporter AcrB transmembrane domain	0.0009 ± 0.0281	1.2 ± 0.3
RF	Multidrug efflux transporter AcrB pore domain; PN1, PN2, PC1 and PC2 subdomains	0.0005 ± 0.0449	1.2 ± 0.4
RF	Multidrug efflux transporter AcrB TolC docking domain; DN and DC subdomains	0.0004 ± 0.0467	1.2 ± 0.4
RF	CBS-domain	0.007 ± 0.032	0.9 ± 0.2
RF	ABC transporter transmembrane region	0.02 ± 0.05	0.7 ± 0.3
RF	NTF2-like	0.00001 ± 0.06132	1.6 ± 0.3
RF	Outer membrane efflux proteins (OEP)	0.0003 ± 0.0323	1.2 ± 0.3
RF	ABC transporter involved in vitamin B12 uptake, BtuC	0.007 ± 0.054	0.8 ± 0.3
RF	Rudiment single hybrid motif	0.008 ± 0.024	0.8 ± 0.2
RF	Mechanosensitive channel protein MscS (YggB),	0.01 ± 0.06	0.6 ± 0.3

	C-terminal domain		
RF	Mechanosensitive channel protein MscS (YggB), transmembrane region	0.02 ± 0.06	0.6 ± 0.2
RF	Proton glutamate symport protein	0.2 ± 0.1	0.3 ± 0.3
RF	Ammonium transporter	0.18 ± 0.07	0.3 ± 0.2
S	Sigma2 domain of RNA polymerase sigma factors	0.0002 ± 0.0290	1.4 ± 0.3
S	ACP-like	0.0001 ± 0.0585	1.3 ± 0.4
S	alpha/beta knot	0.6 ± 0.3	0.3 ± 0.1
S	E set domains	0.002 ± 0.063	1.0 ± 0.3
S	MOP-like	0.001 ± 0.063	1.0 ± 0.3
S	PIN domain-like	0.009 ± 0.045	0.7 ± 0.3
S	Anti-sigma factor antagonist SpoIIaa	0.0006 ± 0.0557	1.1 ± 0.3
S	YjgF-like	0.00008 ± 0.02611	1.3 ± 0.2
S	HCP-like	0.1 ± 0.2	0.4 ± 0.4
S	ITPase-like	0.08 ± 0.07	0.4 ± 0.1
S	MoaD/ThiS	0.04 ± 0.04	0.5 ± 0.2
S	YbaK/ProRS associated domain	0.02 ± 0.05	0.6 ± 0.2
S	Sporulation related repeat	0.1 ± 0.2	0.3 ± 0.2
S	GatB/YqeY motif	0.19 ± 0.09	0.3 ± 0.1
SB	AhpD-like	0.00002 ± 0.03103	1.5 ± 0.3
T	CheY-like	0.00003 ± 0.02882	1.7 ± 0.2
T	PYP-like sensor domain (PAS domain)	0.000002 ± 0.052582	2.0 ± 0.5
T	Homodimeric domain of signal transducing histidine kinase	0.00003 ± 0.02859	1.6 ± 0.3
T	Nucleotide cyclase	0.00002 ± 0.02994	1.6 ± 0.4
T	GAF domain-like	0.00001 ± 0.03280	1.7 ± 0.3
T	PDZ domain-like	0.10 ± 0.08	0.5 ± 0.2
T	EAL domain-like	0.003 ± 0.084	0.9 ± 0.4
T	cAMP-binding domain-like	0.00007 ± 0.01964	1.4 ± 0.3
T	Histidine-containing phosphotransfer domain, HPT domain	0.0001 ± 0.0394	1.2 ± 0.3
T	GlnB-like	0.02 ± 0.09	0.6 ± 0.2
T	Mss4-like	0.01 ± 0.06	0.6 ± 0.3
TA	Sigma3 and sigma4 domains of RNA polymerase sigma factors	0.001 ± 0.042	1.1 ± 0.2
TA	OsmC-like	0.0008 ± 0.0295	1.0 ± 0.2
TA	CinA-like	0.14 ± 0.07	0.3 ± 0.1

Table A.3: Scaling exponent of Pfam clans. The abundance of a clan scales as a power law of the genome size with family-dependent scaling exponents β_i. Each row of the table corresponds to a clan and shows its scaling exponent along with its error and the corresponding functional category (category code). Clans associated to the same functional category are ordered in decreasing order of abundance. The error associated with the exponent is calculated as the root mean square deviation of the logarithm of the category abundance across all genomes from the estimated scaling law.

cat. code	clan name	$A_i \pm \sigma_{A_i}$	$\beta_i \pm \sigma_{\beta_i}$
A	S4 domain superfamily	0.5 ± 0.3	0.29 ± 0.15
C	Pyruvate kinase-like TIM barrel superfamily	0.0005 ± 0.0298	1.16 ± 0.19
C	6-phosphogluconate dehydrogenase C-terminal-like superfamily	0.0007 ± 0.0163	1.10 ± 0.16
C	SIS domain fold	0.01 ± 0.06	0.76 ± 0.21
C	Transmembrane di-heme cytochrome superfamily	0.007 ± 0.056	0.82 ± 0.24
C	Enolase like TIM barrel	0.0003 ± 0.0474	1.13 ± 0.28
C	PFK-like superfamily	0.05 ± 0.06	0.50 ± 0.23
C	LeuD/IlvD-like	0.03 ± 0.03	0.52 ± 0.16
CA	Cytochrome c superfamily	0.005 ± 0.067	0.88 ± 0.45
CA	Acyl-CoA dehydrogenase, C-terminal domain-like	0.00 ± 0.02	1.95 ± 0.42
CA	Rieske-like iron-sulphur domain	0.0002 ± 0.05	1.15 ± 0.30
CA	FMN-dependent nitroreductase-like	0.005 ± 0.053	0.78 ± 0.24
CB	PRC-barrel like superfamily	0.01 ± 0.08	0.57 ± 0.28
E	ACT-like domain	0.03 ± 0.12	0.70 ± 0.20
E	gamma-glutamylcysteine synthetase/glutamine synthetase clan	0.002 ± 0.026	0.91 ± 0.23
E	DAP epimerase superfamily	0.04 ± 0.09	0.51 ± 0.17
E	Arginase/deacetylase superfamily	0.005 ± 0.056	0.74 ± 0.24
E	Aspartate/glutamate racemase superfamily	0.004 ± 0.063	0.74 ± 0.23
F	Ribonuclease H-like superfamily	0.031 ± 0.047	0.78 ± 0.29
F	Nucleotidyltransferase superfamily	0.02 ± 0.06	0.73 ± 0.19
F	PRPP synthetase-associated protein 1	0.2 ± 0.2	0.43 ± 0.15
F	Tetrahydrobiopterin biosynthesis-like enzyme superfamily	0.10 ± 0.09	0.43 ± 0.20
F	Nucleotidyltransferase substrate binding domain	0.1 ± 0.1	0.37 ± 0.26
F	Purine and uridine phosphorylase superfamily	0.07 ± 0.06	0.44 ± 0.21
F	dUTPase like superfamily	0.24 ± 0.07	0.22 ± 0.15
G	Tim barrel glycosyl hydrolase superfamily	0.0002 ± 0.0523	1.32 ± 0.42
G	Six-hairpin glycosidase superfamily	0.0004 ± 0.0425	1.13 ± 0.36
G	Galactose-binding domain-like superfamily	0.003 ± 0.112	0.84 ± 0.46
G	inositol polyphosphate 1 phosphatase like superfamily	0.01 ± 0.06	0.65 ± 0.24
G	HIT superfamily	0.02 ± 0.08	0.55 ± 0.23
GA	Glycosyl transferase clan GT-B	0.008 ± 0.036	0.98 ± 0.19

GA	Pectate lyase-like beta helix	0.00002 ± 0.0538	1.49 ± 0.52
GA	Glycosyl hydrolase domain superfamily	0.003 ± 0.060	0.87 ± 0.26
GA	Double Psi beta barrel glucanase	0.02 ± 0.04	0.55 ± 0.19
H	ATP-grasp superfamily	0.008 ± 0.048	0.88 ± 0.15
H	Acyl-coenzyme A oxidase/dehydrogenase N-terminal	0.000001 ± 0.024165	1.86 ± 0.42
H	Riboflavin synthase/Ferredoxin reductase FAD binding domain	0.003 ± 0.055	0.90 ± 0.23
H	FMN-binding split barrel superfamily	0.00001 ± 0.02020	1.52 ± 0.27
H	Dihydrofolate reductase-like	0.01 ± 0.05	0.65 ± 0.23
H	Release factor superfamily	0.21 ± 0.07	0.29 ± 0.09
H	Succinyl-CoA synthetase flavodoxin domain superfamily	0.16 ± 0.08	0.34 ± 0.18
HA	P-loop containing nucleoside triphosphate hydrolase superfamily	0.7 ± 0.2	0.70 ± 0.07
HA	PCMH-like FAD binding	0.00005 ± 0.0249	1.37 ± 0.25
HD	PhoU-like superfamily	0.09 ± 0.12	0.41 ± 0.20
HE	Ubiquitin superfamily	0.0004 ± 0.0341	1.24 ± 0.31
I	HotDog superfamily	0.00002 ± 0.01312	1.60 ± 0.24
I	Creatinase/prolidase N-terminal domain superfamily	0.02 ± 0.05	0.54 ± 0.20
IA	PLC-like phosphodiesterases	0.02 ± 0.06	0.53 ± 0.23
J	Ribosomal protein S5 domain 2-like superfamily	1.05 ± 3.94	0.32 ± 0.08
J	Transcription elongation factor G C-terminal	0.56 ± 0.2	0.28 ± 0.11
J	Helix-two-turns-helix superfamily	0.5 ± 0.1	0.20 ± 0.17
J	DALR superfamily	0.5 ± 0.2	0.20 ± 0.12
K	Peptidase clan SF	0.01 ± 0.05	0.71 ± 0.20
L	OB fold	0.9 ± 0.8	0.45 ± 0.07
L	PD-(D/E)XK nuclease superfamily	0.1 ± 0.1	0.58 ± 0.23
L	NUDIX superfamily	0.0004 ± 0.0212	1.19 ± 0.21
L	DNA breaking-rejoining enzyme superfamily	0.002 ± 0.022	0.97 ± 0.25
L	His-Me finger endonuclease superfamily	0.005 ± 0.038	0.78 ± 0.28
L	DNase I-like	0.006 ± 0.047	0.75 ± 0.24
L	GIY-YIG endonuclease superfamily	0.10 ± 0.07	0.36 ± 0.22
L	DNA/RNA ligase superfamily	0.02 ± 0.09	0.55 ± 0.20
L	HRDC-like superfamily	0.07 ± 0.06	0.36 ± 0.16
LA	Helix-turn-helix clan	0.0002 ± 0.0216	1.64 ± 0.14
LA	Periplasmic binding protein like	0.00004 ± 0.02710	1.49 ± 0.38
LA	Fatty acid responsive transcription factor FadR, C-terminal domain	0 ± 0.03	1.98 ± 0.34
LA	lambda integrase N-terminal domain	0.06 ± 0.05	0.49 ± 0.22
LA	MetJ/Arc repressor superfamily	0.03 ± 0.10	0.56 ± 0.34
LA	ParB-like superfamily	0.008 ± 0.030	0.70 ± 0.26

LA	IHF-like DNA-binding protein supewrfamily	0.08 ± 0.13	0.39 ± 0.28
LB	EPT/RTPC-like superfamily	0.1 ± 0.1	0.33 ± 0.13
MA	Ig-like fold superfamily (E-set)	0.00004 ± 0.02909	1.49 ± 0.43
MA	von Willebrand factor type A	0.0003 ± 0.0271	1.17 ± 0.30
MA	Pilus subunit	0.005 ± 0.084	0.83 ± 0.41
MA	PGBD superfamily	0.005 ± 0.061	0.75 ± 0.33
MA	Peptidase MD	0.04 ± 0.06	0.48 ± 0.24
N	Flagellar motor switch family	0.1 ± 0.2	0.34 ± 0.30
O	GroES-like superfamily	0.00002 ± 0.02079	1.56 ± 0.28
O	FKBP-like superfamily	0.02 ± 0.04	0.62 ± 0.28
O	Chaperone J-domain superfamily	0.07 ± 0.08	0.45 ± 0.24
O	Cyclophilin-like superfamily	0.006 ± 0.024	0.74 ± 0.19
O	HSP20-like chaperone superfamily	0.008 ± 0.048	0.65 ± 0.24
OA	Peptidase clan MA	0.005 ± 0.031	0.97 ± 0.15
OA	ClpP/Crotonase superfamily	0.002 ± 0.047	1.06 ± 0.21
OA	Peptidase clan MH/MC/MF	0.002 ± 0.033	1.03 ± 0.19
OA	Calcineurin-like phosphoesterase superfamily	0.002 ± 0.023	1.00 ± 0.18
OA	Peptidase clan CA	0.0008 ± 0.0260	1.09 ± 0.24
OA	LuxS/MPP-like metallohydrolase	0.09 ± 0.10	0.48 ± 0.26
OA	Peptidase clan PA	0.005 ± 0.078	0.79 ± 0.26
OA	MACRO domain superfamily	0.10 ± 0.07	0.33 ± 0.18
OB	PP2C-like superfamily	0.0008 ± 0.0557	0.97 ± 0.34
P	Ferritin-like Superfamily	0.001 ± 0.022	1.05 ± 0.20
P	Multicopper oxidase-like domain	0.002 ± 0.046	0.93 ± 0.30
P	SPFH superfamily	0.2 ± 0.2	0.34 ± 0.23
P	SufE/NifU superfamily	0.2 ± 0.1	0.23 ± 0.16
Q	Dimeric alpha/beta barrel superfamily	0 ± 0.02	2.01 ± 0.29
R	Bet V 1 like	0.00002 ± 0.03971	1.50 ± 0.37
R	Acetyl-decarboxylase like superfamily	0.0005 ± 0.0511	1.06 ± 0.29
R	Helical backbone metal receptor superfamily	0.004 ± 0.068	0.81 ± 0.34
R	GME superfamily	0.04 ± 0.09	0.44 ± 0.21
RA	4Fe-4S ferredoxins	0.008 ± 0.133	0.94 ± 0.35
RA	Thioredoxin-like	0.002 ± 0.026	1.16 ± 0.20
RA	VOC superfamily	0.00 ± 0.01	2.11 ± 0.32
RA	Metallo-hydrolase/oxidoreductase superfamily	0.0008 ± 0.0236	1.13 ± 0.17
RA	ALDH-like superfamily	0.000009 ± 0.009462	1.64 ± 0.25
RA	2Fe-2S iron-sulfur cluster binding domain	0.001 ± 0.060	1.05 ± 0.26
RA	Transthyretin superfamily	0.002 ± 0.074	0.94 ± 0.53
RA	Flavoprotein	0.003 ± 0.049	0.89 ± 0.27

RA	Isocitrate/Isopropylmalate dehydrogenase-like superfamily	0.03 ± 0.04	0.64 ± 0.17
RA	Formate/glycerate dehydrogenase catalytic domain-like superfamily	0.001 ± 0.030	0.99 ± 0.19
RA	Ferredoxin / Ferric reductase-like NAD binding	0.0007 ± 0.0564	1.02 ± 0.26
RA	Dehydroquinate synthase-like superfamily	0.01 ± 0.05	0.68 ± 0.29
RA	FAD-linked oxidase C-terminal domain superfamily	0.0001 ± 0.0283	1.21 ± 0.25
RA	Acid phosphatase/Vanadium-dependent haloperoxidase	0.004 ± 0.050	0.77 ± 0.24
RA	LDH C-terminal domain-like superfamily	0.24 ± 0.10	0.23 ± 0.25
RA	FAD-linked oxidoreductase	0.02 ± 0.04	0.52 ± 0.16
RB	PLP dependent aminotransferase superfamily	0.001 ± 0.040	1.22 ± 0.13
RB	N-acetyltransferase like	0.00001 ± 0.02720	1.70 ± 0.23
RB	Glycosyl transferase clan GT-A	0.01 ± 0.06	0.91 ± 0.20
RB	Class-I Glutamine amidotransferase superfamily	0.003 ± 0.032	0.99 ± 0.14
RB	Isomerase,CoA transferase & Translation initiation factor Superfamily	0.001 ± 0.026	1.04 ± 0.22
RB	CoA-dependent acyltransferase superfamily	0.001 ± 0.098	0.96 ± 0.39
RB	Patatin/FabD/lysophospholipase-like superfamily	0.0003 ± 0.0525	1.12 ± 0.26
RB	Acyltransferase clan	0.02 ± 0.08	0.67 ± 0.28
RC	FAD/NAD(P)-binding Rossmann fold Superfamily	0.02 ± 0.01	1.12 ± 0.08
RC	Alpha/Beta hydrolase fold	0.00007 ± 0.03443	1.54 ± 0.29
RC	Actin-like ATPase Superfamily	0.04 ± 0.03	0.75 ± 0.14
RC	Thiolase-like Superfamily	0.0006 ± 0.0464	1.22 ± 0.25
RC	HAD superfamily	0.03 ± 0.10	0.79 ± 0.19
RC	Amidohydrolase superfamily	0.002 ± 0.039	1.07 ± 0.17
RC	Hexapeptide repeat superfamily	0.13 ± 0.05	0.58 ± 0.19
RC	ANL superfamily	0.000001 ± 0.013365	1.91 ± 0.30
RC	Serine beta-lactamase-like superfamily	0.006 ± 0.033	0.90 ± 0.18
RC	HD/PDEase superfamily	0.06 ± 0.12	0.60 ± 0.27
RC	NTN hydrolase superfamily	0.001 ± 0.036	1.06 ± 0.17
RC	Ribokinase-like superfamily	0.005 ± 0.054	0.87 ± 0.20
RC	Alkaline phosphatase-like	0.001 ± 0.039	1.01 ± 0.30
RC	Lysozyme-like superfamily	0.003 ± 0.032	0.89 ± 0.26
RC	LysM-like domain	0.06 ± 0.16	0.53 ± 0.34
RC	DHS-like NAD/FAD-binding domain	0.0002 ± 0.0101	1.20 ± 0.18
RC	Cytidine deaminase-like (CDA) superfamily	0.03 ± 0.07	0.60 ± 0.14
RC	Phospholipase D superfamily	0.04 ± 0.11	0.55 ± 0.28
RC	Histidine phosphatase superfamily	0.006 ± 0.062	0.77 ± 0.26
RC	Glycoside hydrolase/deacetylase superfamily	0.0008 ± 0.0250	1.01 ± 0.25
RC	Galactose Mutarotase-like superfamily	0.002 ± 0.038	0.86 ± 0.31

RC	SGNH hydrolase superfamily	0.002 ± 0.058	0.91 ± 0.32
RC	PFL-like glycyl radical enzyme superfamily	0.3 ± 0.3	0.26 ± 0.31
RC	Enolase N-terminal domain-like superfamily	0.0003 ± 0.0477	1.06 ± 0.27
RC	Fumarylacetoacetate hydrolase, C-terminal domain, superfamily	0.00003 ± 0.02993	1.34 ± 0.29
RC	L,D-transpeptidase catalytic domain	0.001 ± 0.030	0.94 ± 0.30
RC	Chorismate lyase/UTRA superfamily	0.003 ± 0.080	0.82 ± 0.31
RC	MoCo carrier protein-like superfamily	0.06 ± 0.06	0.44 ± 0.16
RC	Chelatase Superfamily	0.04 ± 0.04	0.45 ± 0.24
RC	Fumarate reductase respiratory complex transmembrane subunits	0.1 ± 0.1	0.31 ± 0.19
RD	Tetratrico peptide repeat superfamily	0.005 ± 0.083	1.07 ± 0.44
RD	Common phosphate binding-site TIM barrel superfamily	0.02 ± 0.06	0.91 ± 0.11
RF	Membrane and transport protein	0.003 ± 0.041	1.01 ± 0.30
RF	ABC transporter membrane domain clan	0.007 ± 0.037	0.84 ± 0.26
RF	NTF2-like superfamily	0.0007 ± 0.1039	1.07 ± 0.36
S	Zinc beta-ribbon	0.06 ± 0.04	0.63 ± 0.20
S	ACP-like superfamily	0.000004 ± 0.062939	1.68 ± 0.41
S	SPOUT Methyltransferase Superfamily	0.5 ± 0.2	0.31 ± 0.11
S	PIN domain superfamily	0.004 ± 0.032	0.87 ± 0.27
S	STAS domain superfamily	0.0002 ± 0.0337	1.17 ± 0.30
S	YjgF-like superfamily	0.00004 ± 0.02870	1.32 ± 0.24
S	Phenylalanine- and lysidine-tRNA synthetase domain superfamily	0.18 ± 0.08	0.26 ± 0.17
S	YqeY-like superfamily	0.17 ± 0.083	0.27 ± 0.14
S	Maf/Ham1 superfamily	0.1 ± 0.1	0.30 ± 0.17
SB	AhpD-like superfamily	0.00005 ± 0.05671	1.31 ± 0.28
ST	Type III antifreeze and spore coat polysaccharide	0.01 ± 0.05	0.62 ± 0.23
T	His Kinase A (phospho-acceptor) domain	0.00003 ± 0.02576	1.72 ± 0.19
T	CheY-like superfamily	0.00002 ± 0.02374	1.72 ± 0.23
T	PAS domain clan	0.000002 ± 0.029792	1.92 ± 0.44
T	Nucleotide cyclase superfamily	0.00001 ± 0.02925	1.63 ± 0.44
T	GAF domain-like	0.00 ± 0.01	2.15 ± 0.29
T	PDZ domain-like peptide-binding superfamily	0.04 ± 0.05	0.59 ± 0.18
T	GlnB-like superfamily	0.008 ± 0.060	0.71 ± 0.25
T	Src homology-3 domain	0.3 ± 0.2	0.25 ± 0.36

Bibliography

[1] E. G. Altmann and M. Gerlach. Statistical laws in linguistics. In *Creativity and Universality in Language*, pages 7–26. Springer, 2016.

[2] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.*, 32(D):226–229, 2004.

[3] A. Angelini, A. Amato, G. Bianconi, B. Bassetti, and M. Cosentino Lagomarsino. Mean-field methods in evolutionary duplication-innovation-loss models for the genome-level repertoire of protein domains. *Phys. Rev. E*, 81(2):021919, Feb 2010.

[4] F. E. Angly, D. Willner, A. Prieto-Davó, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, et al. The gaas metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS computational biology*, 5(12):e1000593, 2009.

[5] S. K. Baek, S. Bernhardsson, and P. Minnhagen. Zipf's law unzipped. *New Journal of Physics*, 13(4):043004, 2011.

[6] A. Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

[7] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[8] L. M. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007.

[9] P. Bork, C. Bowler, C. De Vargas, G. Gorsky, E. Karsenti, and P. Wincker. Tara oceans studies plankton at planetary scale. *Science*, 348(6237):873–873, 2015.

[10] C. I. Branden et al. *Introduction to protein structure*. Garland Science, 1999.

[11] V. Charoensawan, D. Wilson, and S. A. Teichmann. Genomic repertoires of dna-binding transcription factors across the tree of life. *Nucleic acids research*, 38(21):7364–7377, 2010.

[12] H. M. P. Consortium et al. Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207, 2012.

[13] O. X. Cordero and P. Hogeweg. Regulome size in prokaryotes: universality and lineage-specific variations. *Trends Genet*, 2009.

[14] B. Corominas-Murtra, R. Hanel, and S. Thurner. Sample space reducing cascading processes produce the full spectrum of scaling exponents.

[15] B. Corominas-Murtra, R. Hanel, and S. Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences*, 112(17):5348–5353, 2015.

[16] M. Cosentino Lagomarsino, A. Sellerio, P. Heijning, and B. Bassetti. Universal features in the genome-level evolution of protein domains. *Genome Biology*, 10(1):R12, 2009.

[17] C. Debes, M. Wang, G. Caetano-Anolles, and F. Graeter. Evolutionary optimization of protein folding. *PLoS Comput Biol*, 9(1):e1002861, 2013.

[18] M. R. Evans and T. Hanney. Nonequilibrium statistical mechanics of the zero-range process and related models. *Journal of Physics A: Mathematical and General*, 38(19):R195, 2005.

[19] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, 2014.

[20] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, et al. Pfam: clans, web tools and services. *Nucleic acids research*, 34(suppl 1):D247–D251, 2006.

[21] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *science*, pages 496–512, 1995.

[22] F. Font-Clos and Á. Corral. Log-log convexity of type-token growth in zipfâĂŹs systems. *Physical review letters*, 114(23):238701, 2015.

[23] M. A. Fortuna, J. A. Bonachela, and S. A. Levin. Evolution of a modular software network. *Proceedings of the National Academy of Sciences*, 108(50):19985–19989, 2011.

[24] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, et al. The minimal gene complement of mycoplasma genitalium. *science*, pages 397–403, 1995.

[25] X. Gabaix. Zipf's law for cities: an explanation. *The Quarterly Journal of Economics*, 114(3):739–767, 1999.

[26] S. J. Giovannoni, J. C. Thrash, and B. Temperton. Implications of streamlining theory for microbial ecology. *The ISME journal*, 8(8):1553, 2014.

[27] J. Gough, K. Karplus, R. Hughey, and C. Chothia. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903 – 919, 2001.

[28] L. Grassi, M. Caselle, M. J. Lercher, and M. Cosentino Lagomarsino. Horizontal gene transfers as metagenomic gene duplications. *Mol Biosyst*, 8(3):790–795, Mar 2012.

[29] L. Grassi, J. Grilli, and M. Cosentino Lagomarsino. Large-scale dynamics of horizontal transfers. *Mob Genet Elements*, 2(3):163–167, May 2012.

[30] J. Grilli, B. Bassetti, S. Maslov, and M. Cosentino Lagomarsino. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucleic Acids Res*, 40(2):530–540, Jan 2012.

[31] J. Grilli, M. Romano, F. Bassetti, and M. Cosentino Lagomarsino. Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers. *Nucleic acids research*, 42(11):6850–60, jun 2014.

[32] R. Hanel and S. Thurner. When do generalized entropies apply? how phase space volume determines entropy. *EPL (Europhysics Letters)*, 96(5):50003, 2011.

[33] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

[34] G. Herdan. *Type-token mathematics*, volume 4. Mouton, 1960.

[35] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, et al. Ebi metagenomicsâĂŤa new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2013.

[36] M. Huynen and E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol*, 15(5):583–589, 1998.

[37] D. Johnston. Stretched exponential relaxation arising from a continuous sum of exponential decays. *Phys. Rev. B*, 74:184430, Nov 2006.

[38] G. P. Karev, Y. I. Wolf, F. S. Berezovskaya, and E. V. Koonin. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol Biol*, 4:32, Sep 2004.

[39] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99, 2013.

[40] D. Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. N. Mantegna, and E. Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PloS one*, 5(12):e15032, 2010.

[41] K. T. Konstantinidis, J. Braff, D. M. Karl, and E. F. DeLong. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station aloha in the north pacific subtropical gyre. *Applied and environmental microbiology*, 75(16):5345–5355, 2009.

[42] E. V. Koonin. Are there laws of genome evolution? *PLoS Comput Biol*, 7(8):e1002173, Aug 2011.

[43] E. V. Koonin. *The logic of chance: the nature and origin of biological evolution*. FT press, 2011.

[44] E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719, 2008.

[45] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218, 2002.

[46] C.-H. Kuo, N. A. Moran, and H. Ochman. The consequences of genetic drift for bacterial genome complexity. *Genome research*, 19(8):1450–1454, 2009.

[47] J. M. Leinaas and J. Myrheim. On the theory of identical particles. *Il Nuovo Cimento B (1971-1996)*, 37(1):1–23, 1977.

[48] R. E. Ley, M. Hamady, C. Lozupone, P. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon. Evolution of mammals and their gut microbes. *Science (New York, N.Y.)*, 320(5883):1647–1651, 06 2008.

[49] A. E. Lobkovsky, Y. I. Wolf, and E. V. Koonin. Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol Evol*, 5(1):233–242, 2013.

[50] N. J. Loman, C. Constantinidou, M. Christner, H. Rohde, J. Z.-M. Chan, J. Quick, J. C. Weir, C. Quince, G. P. Smith, J. R. Betley, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic escherichia coli o104: H4. *Jama*, 309(14):1502–1510, 2013.

[51] R. Louf and M. Barthelemy. Modeling the polycentric transition of cities. *Physical review letters*, 111(19):198702, 2013.

[52] C. Lozupone and R. Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 12 2005.

[53] M. Madan Babu and S. Teichmann. Evolution of transcription factors and the gene regulatory network in escherichia coli. *Nucleic Acids Res*, 31:1234–1244, 2003.

[54] M. A. Mahowald, F. E. Rey, H. Seedorf, P. J. Turnbaugh, R. S. Fulton, A. Wollam, N. Shah, C. Wang, V. Magrini, R. K. Wilson, et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proceedings of the National Academy of Sciences*, 106(14):5859–5864, 2009.

[55] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc Natl Acad Sci U S A*, 106(24):9743–9748, Jun 2009.

[56] B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, J. H. Badger, et al. A framework for human microbiome research. *nature*, 486(7402):215, 2012.

[57] N. Molina and E. van Nimwegen. The evolution of domain-content in bacterial genomes. *Biology Direct*, 3(1):51, 2008.

[58] N. Molina and E. van Nimwegen. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends in genetics : TIG*, 25(6):243–7, jun 2009.

[59] S. Nayfach and K. S. Pollard. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome biology*, 16(1):51, 2015.

[60] S. Nayfach and K. S. Pollard. Toward accurate and quantitative comparative metagenomics. *Cell*, 166(5):1103–1116, 2016.

[61] A. Ndhlovu, P. M. Durand, and S. Hazelhurst. Evodb: a database of evolutionary rate profiles, associated protein domains and phylogenetic trees for pfam-a. *Database*, 2015:bav065, 2015.

[62] C. A. Orengo, A. Michie, S. Jones, D. T. Jones, M. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[63] C. A. Orengo and J. M. Thornton. Protein families and their evolutionâĂŤa structural perspective. *Annual Review of Biochemistry*, 74(1):867–900, 2005.

[64] T. Y. Pang and S. Maslov. A toolbox model of evolution of metabolic pathways on networks of arbitrary topology. *PLoS Comput Biol*, 7(5):e1001137, May 2011.

[65] T. Y. Pang and S. Maslov. Universal distribution of component frequencies in biological and technological systems. *Proc Natl Acad Sci U S A*, 110(15):6235–6239, Apr 2013.

[66] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

[67] D. M. Powers. Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics, 1998.

[68] T. Prakash and T. D. Taylor. Functional assignment of metagenomic data: challenges and applications. *Briefings in bioinformatics*, 13(6):711–727, 2012.

[69] J. Qian, N. M. Luscombe, and M. Gerstein. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol*, 313(4):673–681, Nov 2001.

[70] J. Raes, J. O. Korbel, M. J. Lercher, C. Von Mering, and P. Bork. Prediction of effective genome size in metagenomic samples. *Genome biology*, 8(1):R10, 2007.

[71] J. A. G. Ranea, D. W. A. Buchan, J. M. Thornton, and C. A. Orengo. Evolution of protein superfamilies and bacterial genome size. *J Mol Biol*, 336(4):871–887, Feb 2004.

[72] J. A. G. Ranea, A. Grant, J. M. Thornton, and C. A. Orengo. Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet*, 21(1):21–25, Jan 2005.

[73] R. F. Schwabe and C. Jobin. The microbiome and cancer. *Nature Reviews Cancer*, 13(11):800–812, 2013.

[74] S. Schwartz, I. Friedberg, I. V. Ivanov, L. A. Davidson, J. S. Goldsby, D. B. Dahl, D. Herman, M. Wang, S. M. Donovan, and R. S. Chapkin. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome biology*, 13(4):r32, 2012.

[75] B. Snel, P. Bork, and M. A. Huynen. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome research*, 12(1):17–25, 2002.

[76] D. Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2006.

[77] S. Suweis, J. Grilli, J. R. Banavar, S. Allesina, and A. Maritan. Effect of localization on the stability of mutualistic ecological networks. *Nature communications*, 6, 2015.

[78] T. Thomas, J. Gilbert, and F. Meyer. Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3, 2012.

[79] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz. The dynamics of correlated novelties. *Scientific reports*, 4:5890, 2014.

[80] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, et al. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480, 2009.

[81] F. Turroni, C. Peano, D. A. Pass, E. Foroni, M. Severgnini, M. J. Claesson, C. Kerr, J. Hourihane, D. Murray, F. Fuligni, et al. Diversity of bifidobacteria within the infant gut microbiota. *PloS one*, 7(5):e36957, 2012.

[82] E. van Nimwegen. Scaling laws in the functional content of genomes. *Trends in Genetics*, 19(9):479 – 484, 2003.

[83] T. Větrovskỳ and P. Baldrian. The variability of the 16s rrna gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one*, 8(2):e57923, 2013.

[84] C. Vogel and C. Chothia. Protein family expansions and biological complexity. *PLoS Comput Biol*, 2(5):e48, 2006.

[85] J. Walter and R. Ley. The human gut microbiome: ecology and recent evolutionary changes. *Annual review of microbiology*, 65:411–429, 2011.

[86] J. C. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. *PLoS Comput Biol*, 6(2):e1000667, 2010.

[87] G. Zipf. K.(1968). the psycho-biology of language: An introduction to dynamic philology, 1935.