# Notes on the evolution and architecture of gene content in prokaryotic genomes[1]

Ludovico Calabrese, Università degli Studi di Milano

2018
December

[1] This is a review of the work that's been done and that is currently being done in the field, original sources are given in the bibliography. The review is not meant to be extensive, rather it singles out a certain approach to the subject that makes use of simple, quantitative models to understand the significance of observed patterns.

# Contents

# Introduction

Experimental and conceptual progress in DNA sequencing and DNA annotation has made possible to describe thousands of prokaryotic genomes at multiple levels of coarse-graining and sparked several lines of research in genomics and beyond.

These notes describe genomes abstractly as sets of different units belonging to different classes. The number of units of a certain genome belonging to a certain class is the abundance of such class in the genome. Each genome will be represented by an abundance vector whose components give the abundance of the many classes in the genome.

To be concrete, we shall call these units genes, so as to think of genomes as made of genes in accordance to the popular picture of genomic material. In practice, though, such units do not have to be genes in the technical sense, i.e., regions of DNA coding for a functional protein. In fact, most often in the literature the unit used is not the gene, but rather the so-called protein domain. Furthermore, in principle such units could even represent bits of DNA completely unrelated to protein coding, such as intergenic DNA or introns. In the appendix, we review the various units used in the literature, their biological significance and the way they relate to one another. In the main notes, we will point out when it's necessary to specify concretely what the units are in order to understand results and distinguish possibilities. Otherwise, we will think of these units in an abstract way.

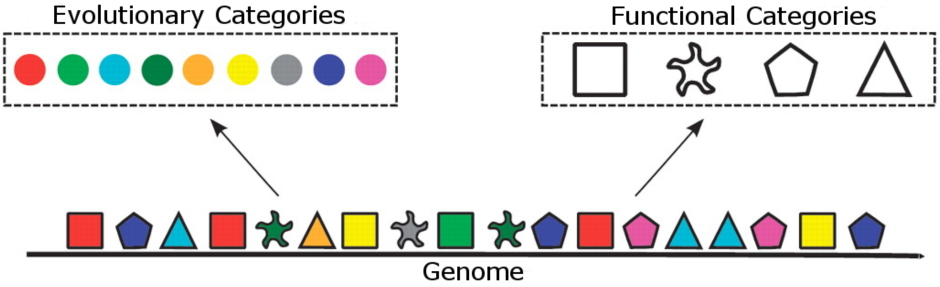

**Figure 1.1: Genomes are made of genes belonging to different classes.**

Therefore, each genome is made of genes and corresponds to an abundance vector representing how many genes in the genome belong to a certain class. Such classes are either evolutionary families, made of genes sharing ancestry, or functional families, made of genes sharing function. Once we consider multiple genomes, we can study the abundance statistics of the entire collection of genes and such study can be taken as the definition of the field of gene content. Numerous patterns have emerged in the literature analysing genomes through this data framework.

The aim of theory in the field of gene content is twofold:

1. to infer the architectural rules governing the gene content structure from the patterns observed analysing the data. One would like to find a 'recipe' by which genomes are built from elementary functional and evolutionary ingredients; for instance the abundances of certain gene families might occur in fixed proportions or the occurrence of certain families might be contingent on the occurrence of other families. A major subgoal of this line of research consist of understanding whether the patterns observed are indeed signatures of design principles or rather simply pervasive statistical patterns.

2. to infer the evolutionary rules that generate genomes from the patterns we observe. In order to figure such rules out, one must think about the physical and biological processes responsible for gene gain and gene loss, make some hypothesis and see if its predictions are consistent with the data. Such work provides insight on various evolutionary mechanisms such as horizontal gene transfer and gene duplication and their relative importance for the whole evolutionary process.

It is apparent that these goals overlaps and results in one direction inform the other and vice versa. The application of these ideas are rather clear as well: in principle we could imagine that only certain abundance patterns allow the realization of biological functionality, so that ultimately certain diseases may be detected by observing deviations from such patterns. This is a refinement on the naive idea according to which 'A person has a gene which causes a certain disease'.[1] The precise description of the various possible patterns requires the development of a quantitative theory and this is the main point of focus of these notes.

[1] Of course biology and medicine have produced all kinds of such refinements since the birth of genomics, coming from different lines of research. Looking at abundance patterns is only one of these possibilities.

# Genome Evolution

## 2.1 A general mathematical framework

Like any evolutionary process, genome evolution can be cast mathematically in terms of a Markov process. The state variable is the abundance vector describing a genome:

$$\mathbf{n} = (n_1, n_2, .., n_F) \tag{2.1}$$

where $F$ is the number of gene families. We shall call $M$ the size of the genome, measured by the total number of genes in the genome:

$$M = \sum_{i=1}^{F} n_i \tag{2.2}$$

Investigating the evolution of genome size is a major research topic by itself, in particular the question of whether natural selection favours bigger or smaller genomes and whether there are biological/physical limit to genome size. The master equation will be:

$$\frac{\partial P(\mathbf{n}, t)}{\partial t} = \sum_{\mathbf{n}'} w(\mathbf{n}|\mathbf{n}')P(\mathbf{n}', t) - w(\mathbf{n}'|\mathbf{n})P(\mathbf{n}, t) \tag{2.3}$$

It's important to understand exactly the biological and physical underpinnings of this equation. The state variable $\mathbf{n}$ represent the genome composition state of a single species or strain. Presumably, the evolution of the genome starts from a common ancestor with $\mathbf{n_0}$, mathematically equivalent to setting $P(\mathbf{n}, 0) = \delta(\mathbf{n} - \mathbf{n_0})$. Transitions represent the evolutionary trajectory of the common ancestor. Each transition involves either the gain or the loss of a gene.

$$\begin{aligned} (n_1, .., n_f, .., n_F) &\to (n_1, .., n_f + 1, .., n_F) \\ (n_1, .., n_f, .., n_F) &\to (n_1, .., n_f - 1, .., n_F) \end{aligned} \tag{2.4}$$

We shall write these transitions in the following manner:

$$\begin{aligned} \mathbf{n} &\to \mathbf{n_f^+} \\ \mathbf{n} &\to \mathbf{n_f^-} \end{aligned} \tag{2.5}$$

so that $\{(\mathbf{n_f^+}, \mathbf{n_f^-})\}_{f=1}^F$ defines all the available neighboring states of $\mathbf{n}$.

The transition rate is usually factored into two contributions, conceptually if not mathematically, following a typical scheme common in population genetics, itself a version of the Darwinian scheme of random variation plus natural selection:

$$w(\mathbf{n}'|\mathbf{n}) = \mu(\mathbf{n}'|\mathbf{n})\phi(\mathbf{n}'|\mathbf{n}) \tag{2.6}$$

$\mu(\mathbf{n}'|\mathbf{n})$ is a mutation rate and it depends on the physical process causing random gene acquisition and deletion. Even if a gene acquisition or deletion event happens, there is no guarantee that natural selection will allow its fixation in the species, i.e., that all member of the species end up with that mutation so that it is really possible to speak of the genome of the species. The second term $\phi(\mathbf{n}'|\mathbf{n})$ models such probability of fixation and it depends on the fitness gain associated with the particular gene acquired or deleted.[1] It's clear that within this framework the entire modelling goes into choosing appropriately the transition rates. We also note that the timescales involved in the equations are obviously evolutionary ones: each transition corresponds to a fixation event which spans a number of generations monotonic in the effective population size, often proportional to it. Multiplying the lifespan of the organism for the effective population size gives a rough estimate of the timescales involved.

In principle the rates $w(\mathbf{n}'|\mathbf{n})$ could depend on time, for instance if transitions between genome composition states depend on the particular environment organisms live in and such environment changes with time. If this was the case, finding the functional form of the rates would be practically impossible. In the following we are going to assume that if rates depend on time, they do so weakly and at any rate we can identify general trends in evolution by removing the time variation of the rates.

Before moving on to specific choices of the rates, we remind that to each Markov process there is a corresponding underlying Markov chain. We write down the equation for such Markov chain because occasionally it's more convenient to use than its continuous-time parent.

$$\Delta P(\mathbf{n}, m) = \sum_{\mathbf{n}'} \pi(\mathbf{n}|\mathbf{n}')P(\mathbf{n}', m) - \pi(\mathbf{n}'|\mathbf{n})P(\mathbf{n}, m) \tag{2.7}$$

[1] If all of this feels somewhat wrong or ridden with hidden assumptions, it's because it is actually: in equation 2.3 we are already making a huge simplifying step because actual evolution acts on the whole population. The real state variable is $\{N(\mathbf{n})\}$, expressing the number of individuals of the population with a certain genome composition vector $\mathbf{n}$. Instead of writing a master equation for these numbers, we are assuming we can instead write an equation for the typical genome composition vector and ignore all variation in the population without losing much. This is sometimes called the monomorphic limit of population genetics and admittedly its validity is by no means obvious.

In the corresponding Markov chain, the evolution of the system is parametrized by the number of transition $m$. We can distinguish transitions in which a gene is gained and transitions in which a gene is lost:

$$m = m_+ + m_- \tag{2.8}$$

Both $m_+$ and $m_-$ are random variables. If the original common ancestor had size $M_0$, then we also have:

$$M - M_0 = m_+ - m_- \tag{2.9}$$

Finally $\pi(\mathbf{n}'|\mathbf{n})$ is now an actually probability instead of a probability rate and it relates to the previous rates in the following way:

$$\pi(\mathbf{n}'|\mathbf{n}) = \frac{w(\mathbf{n}'|\mathbf{n})}{\sum_{\mathbf{n}'\neq\mathbf{n}} w(\mathbf{n}'|\mathbf{n})} = \frac{w(\mathbf{n}'|\mathbf{n})}{w(\mathbf{n})} = \frac{w(\mathbf{n}'|\mathbf{n})}{1 - \frac{1}{\tau(\mathbf{n})}} \tag{2.10}$$

where $w(\mathbf{n})$ is the total rate out of state $\mathbf{n}$, while $\tau(\mathbf{n})$ is the typical time spent in state $\mathbf{n}$.

Note that both the distribution of genome sizes $M$ and the distribution of genome composition states $\mathbf{n}$ are outcomes of the evolutionary process and therefore any true model of genome evolution must be able at least in principle to generate both such observables. On other hand, it is both interesting and helpful to think about these observable separately, i.e., mathematically:

$$P(\mathbf{n}, t) = \sum_M P(\mathbf{n}|M, t)P(M, t) \tag{2.11}$$

## 2.2 Evolution of genome size

The evolution of genome size is a major topic by itself and in addition, it affects directly evolutionary models of genome composition state. For instance when considering the transition rate $w(\mathbf{n_f^+}|\mathbf{n})$, we might model it by considering the probability rate of gaining a gene, no matter which, times the probability that such gene belong to family $f$, effectively writing $w(\mathbf{n_f^+}|\mathbf{n}) = g(M)P_f(\mathbf{n})$.[2] The form of such $g(M)$, the probability rate of gaining a gene given the genome current size $M$, affects the whole genome composition dynamics, while also being the natural input of a Markov process describing exclusively genome size (as $g(M) = g(M + 1|M)$).

[2] This is not done without making some basic assumptions. In the Markov chain formalism, the probability of gaining a gene starting from state $\mathbf{n}$ is $\sum_f \pi(\mathbf{n_f^+}|\mathbf{n}) = G(\mathbf{n}) = G(n_1, .., n_f)$. The basic assumption then is $G(n_1, .., n_f) = G(\sum_f n_f) = G(M)$.

### 2.2.1 Master Equation

We write a master equation for the evolution of genome size in the Markov chain formalism. At each step, the genome either gains or loses a gene, so that the probability of gain and loss are simply related by normalization in this case. $L(M)$ and $G(M)$ are the probabilities of loss and gain at fixed size.

$$\Delta P(M) = L(M{+}1)P(M{+}1) - G(M)P(M) + G(M{-}1)P(M{-}1) - L(M)P(M) \tag{2.12}$$

To make the notation less heavy, we dropped the time parameter $m$, the total number of transitions, but it's obviously implied. After taking into account normalization:

$$\Delta P(M) = G(M-1)P(M-1) - G(M+1)P(M+1) + P(M+1) - P(M) \tag{2.13}$$

This is essentially a random walk in 1-D with the added presence of a non-uniform field in the one-dimensional strip.

### 2.2.2 Existence of a steady state

We are particularly interested in the existence and form of a steady state. To this end, we set the left-hand side of the previous equation to zero and make use of a diffusion limit on the right-hand side. We obtain the following equation:

$$\frac{\partial}{\partial M}[1 - 2G(M)]P_{ss}(M) + \frac{1}{2}\frac{\partial^2}{\partial M^2}P_{ss}(M) = 0 \tag{2.14}$$

which translates to

$$[1 - 2G(M)]P_{ss}(M) + \frac{1}{2}\frac{\partial}{\partial M}P_{ss}(M) = c \tag{2.15}$$

The normalization condition $\int_0^M P_{ss}(M')dM' = 1$ and a necessary implication of this fact is $P_{ss}(M) \to 0$ as $M \to \infty$. Then $c = 0$ and we can write:

$$\frac{\partial}{\partial M}P_{ss}(M) = -2[1 - 2G(M)]P_{ss}(M) \tag{2.16}$$

Admittedly, there is a slight technical complication. Equation 2.13 does not hold for $M = 0$ (biologically we expect it to be wrong much sooner in fact). The state $M = 0$ works as an absorbing state, breaking the ergodicity of the Markov chain and making the existence of a unique steady state impossible. The real normalization condition therefore is:

$$P_{ext}(M_0) + \int_0^\infty P_{ss}(M')dM' = 1 \qquad (2.17)$$

The way out of this is to consider $M_0 \gg 0$ where $P_{ext}(M_0) \to 0$ and we can disregard the absorbing state. Actually, in order to do this we also need some conditions on $G(M)$: for instance if loss was always favourable, it wouldn't even matter where we started, since at sufficiently long times we'd reach the absorbing state anyway. On the other hand, the equilibrium state is trivial in this case as the system will necessarily reach the absorbing state, i.e., lose as many genes as possible until it reaches the minimal size for a functional genome. Although this is possible in certain environmental niches, such situation does not seem to capture overall trends in bacterial genome sizes as in reality they span several orders of magnitude.

We can then integrate equation 2.16 and apply the normalization condition. One obtains:

$$P_{ss}(M) = \frac{\exp\left(-2\int_0^M [1-2G(M')]dM'\right)}{\int_0^\infty \exp\left(-2\int_0^M [1-2G(M')]dM'\right)dM} \qquad (2.18)$$

This is a true steady state only if the normalization condition holds. To this end, we need to study what of $G(M)$ allows normalization. The normalization condition is equivalent to:

$$\int_0^\infty \exp\left(-2\int_0^M [1-2G(M')]dM'\right)dM < \infty \qquad (2.19)$$

As $0 < G(M) < 1$, then $-1 < 1 - 2G(M) < 1$. The sign of $1 - 2G(M)$ tells us whether it is a gene gain or gain loss that is favourable at size $M$. It's positive when loss is favourable and negative when gain is favourable.

To gain intuition, we shall consider some specific forms of $G(M)$. Let us call $1 - 2G(M)$ with $h(M)$. The most natural way to think about the equilibrium situation is to pick $G(M)$ such that $\exists M^* : \forall M > M^*\, h(M) > 0 \wedge \forall M < M^*\, h(M) > 0$. Above such $M^*$ the system tends to lose genes, while below it tends to gains genes. $h(M)$ is a kind of sigmoid function and it's natural to think of $M^*$ as some mean equilibrium value around which the system fluctuates. Strictly speaking, for the normalization condition to occur what's important is only half of the condition, i.e., $\exists M^* : \forall M > M^*\, h(M) > 0$ so that it is not possible for the system to grow indefinitely. On the other hand, if $h(M) > 0\, \forall M$ the equilibrium state is trivial as described before and therefore we are interested in situations with regions $h(M) < 0$.

### 2.2.3 An analytically solvable example

For concreteness, we shall choose a particular sigmoid function:

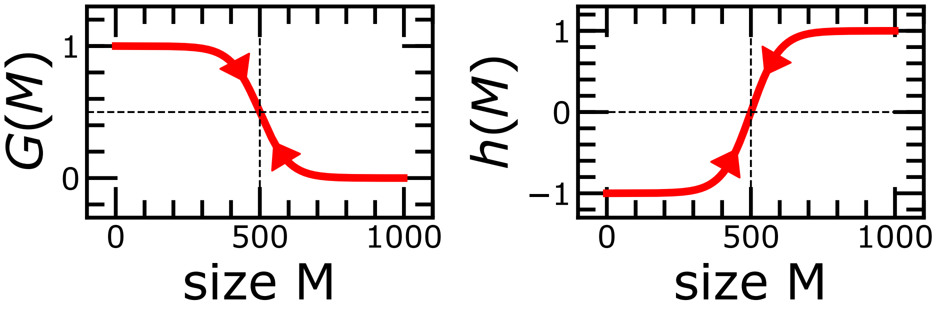

**Figure 2.1: The prototypical shape of $\mathbf{G(M)}$ and $\mathbf{h(m) = 1 - 2G(M)}$ allowing the existence of a steady state.** The inflection point in both plots gives the equilibrium value of the genome size. The arrows indicate the tendency of the system to gain or lose genes before and after the equilibrium value.

$$h(M) = \tanh\left[k(M - M^*)\right] \tag{2.20}$$

$M^*$ is the equilibrium, while $k$ is parameter that describes how fast the system changes regime near the equilibrium, i.e., how quickly gene loss becomes favourable over gene gain. It is related to equilibrium fluctuation as we will see in the following.

Let us define:

$$H(M) = 2\int_0^M h(M')dM' \tag{2.21}$$

so that $P_{ss} \propto \exp\left[-H(M)\right]$. This choice allows easy analytical computations to derive the entire size distribution. It is easy to see that:

$$H(M) = \frac{2}{k}\ln\left[\frac{\cosh k(M - M^*)}{\cosh kM^*}\right] = \frac{2}{k}\ln\left[\cosh k(M - M^*)\right] + H(M^*) \tag{2.22}$$

and therefore we obtain for the distribution:

$$\frac{P_{ss}(M)}{P_{ss}(M^*)} = \exp\left\{-\left[H(M) - H(M^*)\right]\right\} = \left[\frac{1}{\cosh k(M - M^*)}\right]^{\frac{2}{k}} \tag{2.23}$$

We shall consider the limits $k(M - M^*) \gg 1$ and $k(M - M^*) \ll 1$.

1. For $k(M-M^*) \gg 1$, $\cosh k(M-M^*) \approx \exp\left[k(M - M^*)\right]$ so we obtain:

$$P_{ss}(M) \propto \exp\left(-(M - M^*)\right) \tag{2.24}$$

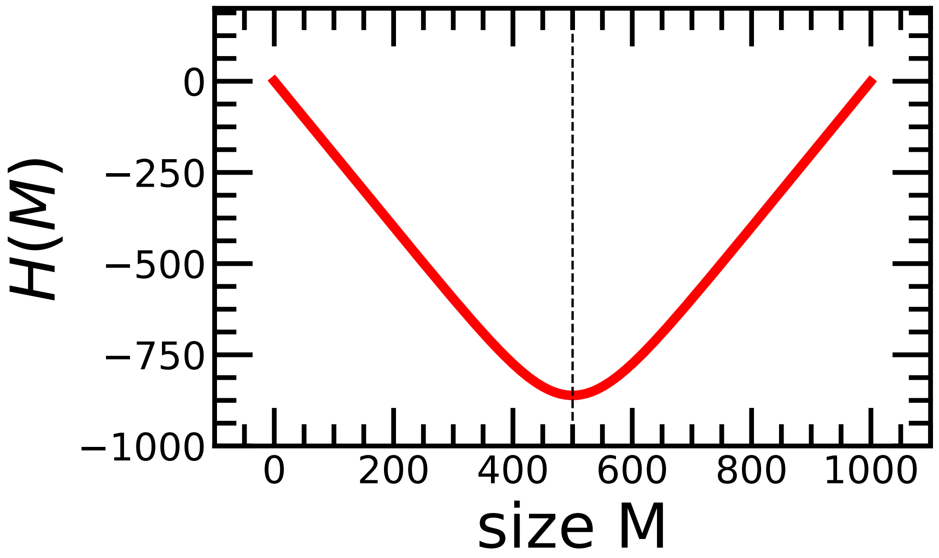

**Figure 2.2: The prototypical form of $\mathbf{H(M) = 2\int_0^M h(M')dM'}$ when steady state is possible.** $H(M)$ presents a clear minimum describing the equilibrium value. $H(M)$ plays exactly the same role of a one-dimensional hamiltonian in statistical physics.

2. For $k(M-M^*) \ll 1$, $\cosh k(M-M^*) \approx 1+\frac{1}{2}\left[k(M-M^*)\right]^2$. It follows that we have a gaussian approximation near $M^*$:

$$P_{ss}(M) \propto \exp\left[-k(M-M^*)^2\right] \tag{2.25}$$

and we can approximate the fluctuations with $\sigma = \sqrt{\frac{1}{2k}}$, which establishes a quantitative relation between $k$ and the size variation.

To conclude, it is easy to imagine modification of the function $h(M)$ encoding different biological situations. Multiple zeroes of this function for instance would imply multiple equilibrium values.

Finally, we also remark that the main use of this model consists of connecting the data to quantities that are harder to access empirically. It may seem underwhelming to obtain essentially a Gaussian distribution in the genome size, but the real value of the model is the connection between the dynamical quantities and the 'static' ones.

### 2.2.4 Relaxation Dynamics in Mean Field

The following equation holds in full generality:

$$\langle M(m+1)\rangle - \langle M(m)\rangle = \langle 2G\left[M(m)\right] - 1\rangle \tag{2.26}$$

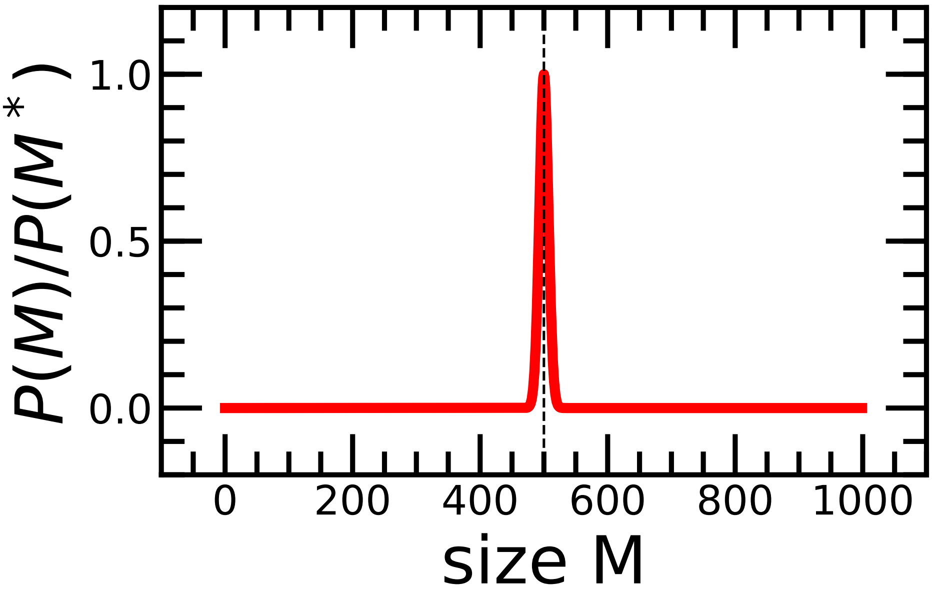

**Figure 2.3: The prototypical form of $\frac{\mathbf{P_{ss}(M)}}{\mathbf{P_{ss}(M^*)}}$ when steady state is possible.** The probability distribution is centered around the equilibrium value $M^*$. Using the concrete expression for $G(M)$ defined in the text, we have used $k = 0.01$ and therefore the typical size of the fluctuations is given by $\sigma = \sqrt{\frac{1}{2k}} \approx 10$, which is what we see in the plot.

In mean-field, it simplifies considerably:

$$\langle \Delta M(m) \rangle = 2G\left[\langle M(m) \rangle\right] - 1 \tag{2.27}$$

We consider a linear version of $G(M)$, equivalent to taking the limit $k(M - M^*) \ll 1$, i.e., relatively big fluctuations.

$$G(M) = \frac{1}{2} - k(M - M^*) \tag{2.28}$$

Then, we have the following dynamic equation in the mean:

$$\langle \Delta M(m) \rangle = -2k(\langle M(m) \rangle - M^*) \tag{2.29}$$

This equation is readily solvable and we obtain:

$$\langle M(m) \rangle = M^* - (1 - 2k)^m (M^* - M_0) \tag{2.30}$$

Thus, the typical timescale to equilibrium, in terms of total number of gene gains and losses, is $\approx |\ln(1 - 2k)|$ and therefore we can connect the fluctuations in steady state to the relaxation dynamic to equilibrium.

We shall now consider a small $k$ limit with respect to time. In the limit $k \ll \frac{1}{m}$, or rather $m \ll \frac{1}{k}$ so that we far from equilibrium, the mean dynamics behaves as follows:

$$\langle M(m) \rangle \approx M_0 + 2km(M^* - M_0) \tag{2.31}$$

which means that $d\langle M(m) \rangle \propto dm$, so the size of the genome can also be used as a time parameter as long as we are far from equilibrium.

### 2.2.5 Models in Real Time

In the following we will consider a series of stylised model in real time, computing the average and the fluctuations.

The average is computed from the following equation:

$$\frac{\partial \langle M(t) \rangle}{\partial t} = \langle g(M) - l(M) \rangle \tag{2.32}$$

The fluctuations require the knowledge of $M(t)^2$. To this end, let us consider $M(t)^2$. $M(t+dt)^2 - M(t)$ is a random variable that can take only value in $\{1 - 2M(t), 0, 1 + 2M(t)\}$ with probabilities $g(M)dt$, $1 - g(M)dt - l(M)dt$ and $l(M)dt$. It is immediate to write

$$\frac{\partial \langle M(t)^2 \rangle}{\partial t} = \langle g(M) + l(M) \rangle + \langle 2M \left[ g(M) - l(M) \right] \rangle \tag{2.33}$$

which is equal to

$$\frac{\partial \langle M(t)^2 \rangle}{\partial t} = \frac{\partial \langle m(t) \rangle}{\partial t} + \langle 2M \left[ g(M) - l(M) \right] \rangle \tag{2.34}$$

This is hard to solve with full generality, but we can start building some intuition by considering some simple cases.

#### 2.2.5.1 Constant Rates

Define $\alpha = g(M) - l(M) = g - l$ and $\beta = g(M) + l(M) = g + l$. One could guess the result already, as this model is simply a series of independent jumps and therefore the central limit theorem holds. At any rate it is easy to prove explicitly:

$$\langle M(t) \rangle = M_0 + \alpha t \tag{2.35}$$

$$\langle M(t) \rangle^2 = M_0^2 + \alpha^2 t^2 + 2M_0 \alpha t \tag{2.36}$$

$$\langle M(t)^2 \rangle = M_0^2 + \beta t + \alpha^2 t^2 + 2M_0 \alpha t \tag{2.37}$$

therefore

$$\mathrm{Var}M(t) = \beta t \tag{2.38}$$

and

$$\frac{\sigma_M(t)}{\langle M(t)\rangle} = \frac{\sqrt{\beta t}}{M_0 + \alpha t} \tag{2.39}$$

which is always less than 1 and strictly decreasing in time.

#### 2.2.5.2 Preferential attachment

Let us write $g(M) = g \cdot M$, $l(M) = l \cdot M$, $\alpha = g - l$.

$$\frac{\partial\langle M(t)\rangle}{\partial t} = \alpha\langle M(t)\rangle \tag{2.40}$$

giving an exponential size increase in time

$$\langle M(t)\rangle = M_0 \exp(\alpha t) \tag{2.41}$$

It follows immediately that it's also true

$$\frac{\partial\langle M(t)^2\rangle}{\partial t} = \beta\langle M(t)\rangle + 2\alpha\langle M(t)^2\rangle \tag{2.42}$$

and

$$\frac{\partial \mathrm{Var}M(t)}{\partial t} = \beta\langle M(t)\rangle + 2\alpha \mathrm{Var}M(t) \tag{2.43}$$

which is solvable. By considering the equation for $\mathrm{Var}M(t) \cdot \exp(-\alpha t)$ we obtain

$$\mathrm{Var}M(t) = M_0\left[\exp(2\alpha t) - \exp(\alpha t)\right] = \langle M(t)\rangle \frac{\beta}{\alpha}\left[\exp(\alpha t) - 1\right] \tag{2.44}$$

Then:

$$\frac{\sigma_M(t)}{\langle M(t)\rangle} = \sqrt{\frac{\exp(\alpha t) - 1}{M_0 \exp(\alpha t)}\frac{\beta}{\alpha}} \approx \sqrt{\frac{\beta}{M_0\alpha}} \tag{2.45}$$

where the last line is obviously true only if $\alpha > 0$.

#### 2.2.5.3 Mixing constant rates and preferential attachment

Let us write $g(M) = g$, $l(M) = l \cdot M$, assuming a constant gain rate and a loss rate increasing with the size of the genome.

$$\frac{\partial \langle M(t) \rangle}{\partial t} = g - l\langle M(t) \rangle \tag{2.46}$$

Then

$$\langle M(t) \rangle = \frac{g}{l} + \left( M_0 - \frac{g}{l} \right) \exp(-l \cdot t) \tag{2.47}$$

Note the presence of a stationary mean value; implicit in all of the above is $g > l$ as otherwise the genome simply loses all the genes. Note that this model is a simple example of the steady state condition analysed previously The equation in $M(t)^2$ is particularly simple near the stationary value of the mean:

$$\frac{\partial \langle M(t)^2 \rangle}{\partial t} = g + l\langle M(t) \rangle + 2g\langle M(t) \rangle - 2l\langle M^2(t) \rangle = 2g + \frac{2g^2}{l} - 2l\langle M^2(t) \rangle \tag{2.48}$$

then

$$\langle M^2(t) \rangle = \frac{g}{l} + \frac{g^2}{l^2} + \left( M_0^2 - \frac{g}{l} - \frac{g^2}{l^2} \right) \exp(-2l \cdot t) \tag{2.49}$$

and effectively

$$\langle M^2(t) \rangle = \frac{g}{l} + \frac{g^2}{l^2} \tag{2.50}$$

implying

$$\mathrm{Var} M(t) = \frac{g}{l} \tag{2.51}$$

and

$$\frac{\sigma_M(t)}{\langle M(t) \rangle} \approx \sqrt{\frac{1}{\langle M(t) \rangle}} = \sqrt{\frac{l}{g}} \tag{2.52}$$

at equilibrium.

## 2.3 Evolution of genome composition: Effective Independent Models

### 2.3.1 Nimwegen's model

We shall begin the description of the full evolution of genome composition by using the insights of Nimwegen. Let us rewrite the general master's equation in 2.3 using the specific transitions defined in 2.5:

$$\frac{\partial P(\mathbf{n},t)}{\partial t} = \sum_{f}^{F} w(\mathbf{n}|\mathbf{n_f^+})P(\mathbf{n_f^+},t)+w(\mathbf{n}|\mathbf{n_f^-})P(\mathbf{n_f^-},t)-\left[w(\mathbf{n_f^+}|\mathbf{n}) + w(\mathbf{n_f^-}|\mathbf{n})\right] P(\mathbf{n},t) \tag{2.53}$$

The rates of the model are defined in the following way:

$$\begin{aligned} w(\mathbf{n_f^+}|\mathbf{n}) = &\quad g(M)P(f) \\ w(\mathbf{n_f^-}|\mathbf{n}) = &\quad l(M)P(f) \end{aligned} \tag{2.54}$$

i.e., the probability rate of gaining/losing a gene belonging to family $f$ is equal to the probability rate of gaining/losing any gene times the probability of such gene belonging to family f. $P(f)$ depends on $\mathbf{n}$ generally. Nimwegen and Molina claim a particular form for this function, in the spirit of abstraction we maintain a completely generic form for now.

Splitting the rate in such way is very convenient mathematically and conceptually, as we will show, but it may raise some highbrows biologically. For instance, causally, genome size surely increases as a consequences of evolutionary pressures on the various families, instead of gene families growing as a consequence of a generic advantage gained with bigger sizes. The equation is not to be read causally, though: it's a simple mathematical rewriting instead as we define $g(M) := \sum_f^F w(\mathbf{n_f^+}|\mathbf{n})$ and $P(f) := \frac{w(\mathbf{n_f^+}|\mathbf{n})}{g(M)}$. As we pointed out before, $\sum_f^F w(\mathbf{n_f^+}|\mathbf{n})$ being dependent only on $M$ is not completely obvious, but a good null assumption nonetheless. Besides this, the main advantage of this writing is conceptual as we separate different processes and it's easier to make hypothesis on the functional forms of $g(m)$ and $P(f)$ than making hypothesis on the whole rate.

Finally, why should $P(f)$ be identical in the gain and loss rate? Nimwegen and Molina assume this symmetry, but admittedly I cannot think of a satisfactory explanation. If anything if a family is more likely to get fixated over another family in the case of a gene addition event, then it seems natural that it is less likely to get fixated in the case of a gene deletion event, as presumably fitness is increased when that family is added and therefore fitness is decreased when that family is removed. Nonetheless, to explore rigorously the model of Nimwegen and for the sake of simplicity, we assume such symmetry and draw its conclusions.

Having defined the model, let us write four equations capturing the average behavior:

$$\begin{aligned}\frac{\partial\langle n_f\rangle}{\partial t} &= \langle [g(M) - l(M)]\, P(f)\rangle \\ \frac{\partial\langle M\rangle}{\partial t} &= \langle g(M) - l(M)\rangle \\ \frac{\partial\langle m_f\rangle}{\partial t} &= \langle [g(M) + l(M)]\, P(f)\rangle \\ \frac{\partial\langle m\rangle}{\partial t} &= \langle g(M) + l(M)\rangle\end{aligned} \tag{2.55}$$

respectively for the number of genes $n_f$ belonging to family $f$, the genome size $M$, the number of transitions/evolutionary events $m_f$ involving family $f$ and the total number of transitions/evolutionary events $m$.

These equations are easy to find considering infinitesimal increments. For example, for sufficiently small $dt$, it is true that $n_f(t+dt)-n_f(t) \in \{-1, 0, 1\}$. Fixing the composition state at $t$, the probabilities of these values are respectively $\{l(M)P(f), g(M)P(f), 1 - l(M)P(f) - g(M)P(f)\}$ which let us obtain the conditional mean. Then, we average over the initial composition state to obtain the global mean.

At this point, it's very tempting to split the average in equations 1 and 3, so that we obtain:

$$\begin{aligned}\frac{\partial\langle n_f\rangle}{\partial t} &= \frac{\partial\langle M\rangle}{\partial t}\,\langle P(f)\rangle \\ \frac{\partial\langle m_f\rangle}{\partial t} &= \frac{\partial\langle m\rangle}{\partial t}\,\langle P(f)\rangle\end{aligned} \tag{2.56}$$

and

$$\begin{aligned}\frac{\partial\langle n_f\rangle}{\partial\langle M\rangle} &= \langle P(f)\rangle \\ \frac{\partial\langle m_f\rangle}{\partial\langle m\rangle} &= \langle P(f)\rangle\end{aligned} \tag{2.57}$$

Is such splitting justified? It certainly is using the mean-field approximation or in simple exact cases, such as a constant gene gain and loss. On the other hand, in general $P(f)$ depend on the size $M$. In a 'neutral' theory for instance, $P(f) = \frac{n_f}{M}$ and technically we cannot write the mean of a multiplication as the multiplication of the means. It remains to be investigated how much higher-order influence the average behavior.

At any rate, even if the splitting isn't possible, it's very convenient to remove time, essentially inaccessible experimentally and consider instead the ratios of the increments. Thanks to this procedure, we can test our theories with respect to $P(f)$ or infer such function from the data. The link between the equations above and the data is not obvious and deserves its own subsection.

#### 2.3.1.1 Connecting the model and the data

Let us review the meaning of equation 2.57.

$$\frac{\partial \langle n_f \rangle}{\partial \langle M \rangle} = \frac{\langle n_f(t+dt)\rangle - \langle n_f(t)\rangle}{\langle M(t+dt)\rangle - \langle M(t)\rangle} = \langle f(\mathbf{n}(t), M(t))\rangle = f(t) \tag{2.58}$$

The second equality stresses the fact that this is actually a function of time only, ultimately, despite the fact that it's often not very instructive to write explicitly the time dependence. To actually measure this ratio, we would need at the very least two snapshots of an evolutionary trajectory. At time $t$ a bunch of genomes are present and we take the average $n_f$ and $M$, then at time $t+dt$ we take once again such averages and finally compute the ratio. This is already impossible, as we observe only genomes at the current time t. Furthermore, the above procedure is likely flawed as the averages in the equation are ensemble averages and the model is not self-averaging, in particular for small $dt$. Therefore, we would need to start at $t_0$ until $t$, take the averages and then actually restart the whole evolutionary process from $t_0$ again until $t+dt$ this time. Replaying the evolutionary tape is even more difficult than having a single complete evolutionary trajectory, so this is obviously not feasible.

On the other hand, we do observe $P(n_f|M,t)$ and its corresponding $\langle n_f(M,t)\rangle$ at our current time. Let us consider the quantity:

$$\frac{\partial \langle n_f(M,t)\rangle}{\partial M} = \frac{\langle n_f(M+\Delta M, t)\rangle - \langle n_f(M,t)\rangle}{\Delta M} \tag{2.59}$$

This quantity is easily accessible in the data and we will see in the next chapter that it's the building block of the stochastic growth models used to investigate the observed genomic architecture.

To relate $\frac{\partial \langle n_f(M,t)\rangle}{\partial M}$ to $\frac{\partial \langle n_f\rangle}{\partial \langle M\rangle}$, we begin by noting:

$$\langle n_f(t)\rangle = \sum_f n_f \sum_M P_t(n_f|M)P_t(M) \approx \sum_f n_f P_t(n_f|\langle M(t)\rangle) = \langle n_f(\langle M(t)\rangle)\rangle_t \tag{2.60}$$

This is a mean-field approximation. Its validity depends on the underlying size distribution and in the appendix we give formal criteria judging the correctness of the approximation. At time $t+dt$

$$\langle n_f(t+dt)\rangle = \langle n_f(\langle M(t+dt)\rangle)\rangle_{t+dt} \tag{2.61}$$

Next, we assume time invariance and therefore

$$\langle n_f(t+dt)\rangle = \langle n_f(\langle M(t+dt)\rangle)\rangle_t \tag{2.62}$$

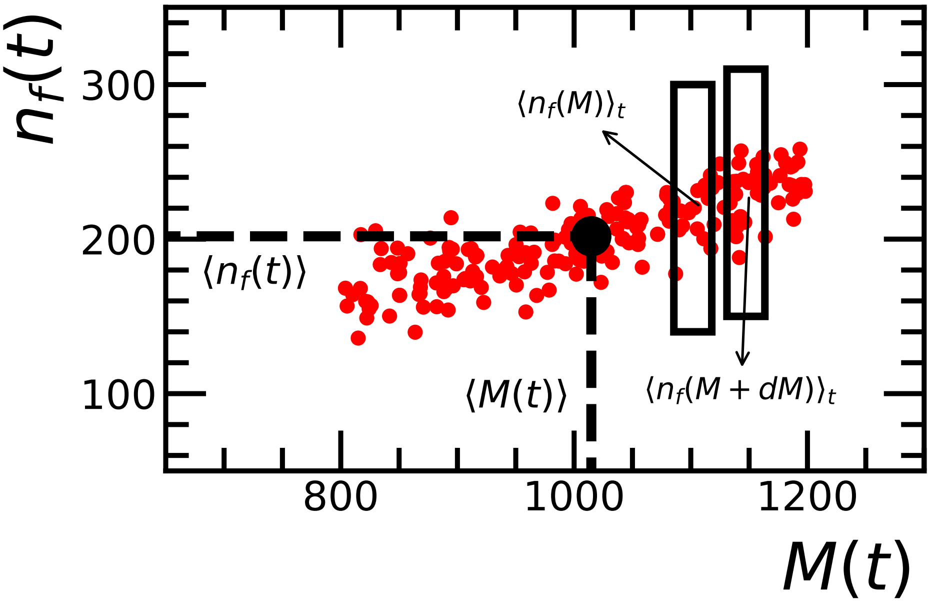

**Figure 2.4: A snapshot of genome evolution at time t.** Each point $(n_f, M)$ represent a genome as seen at time $t$. The big black point represent the average genome at time $t$. Taking another snapshot at time $t + dt$ would see the cloud of points move, in particular the motion of the average genome is easily described by the model. The points inside the black boxes define the conditional probabilities $P_t(n_f, M)$ and their respective averages $\langle n_f(M)\rangle_t$. In the text we prove that, by assuming time invariance, the way the average genome $(\langle n_f(t)\rangle, \langle M(t)\rangle)$ moves with time is tightly related to how gene families scale with size at any fixed $t$.

We keep the subscript $t$ in the conditional average to remind us that such averages are obtained through data sampled at the current time, but there is no real time dependence.

Then:

$$\langle n_f(t+dt)\rangle - \langle n_f(t)\rangle = \langle n_f(\langle M(t)\rangle + \Delta M(t)\rangle_t - \langle n_f(\langle M(t)\rangle)\rangle_t \tag{2.63}$$

Let now $t$ such that $\langle M(t)\rangle = M$. It's easy to obtain:

$$\frac{\langle n_f(t+dt)\rangle - \langle n_f(t)\rangle}{\langle \Delta M(t)\rangle} = \frac{\partial \langle n_f(M)\rangle_t}{\partial M} \tag{2.64}$$

The left hand-side is the function $P(f)$ defined by the model with the proper change of variables from $t$ to $M$ defined by the previous equation. Therefore $P(f)$ is accessible through the data by looking at the curves $\langle n_f(M)\rangle_t$.

### 2.3.2 Birth, death and innovation model

## 2.4 Evolution of genome composition: Interacting Models

## 2.5 Refining the Markovian Framework: the Starwars Effect

A major factual mistake of the general mathematical framework outlined 2.1 is the assumption of genome independence: all the genomes are generated by independent runs of the evolutionary process. This is quite obviously not the case. For one thing, all genome descend from a single common ancestor. This is actually taken into account by the various models described in the previous sections by considering a common initial condition. More importantly, macroevolution proceeds by speciation, so genomes don't just share a single common ancestor, but different groups of genomes have their common ancestors, themselves ultimately coming from the LUCA. The $N$ genomes observed now are not simply $N$ independent iterations of the same process coming from the LUCA, but instead they share entire portions of their evolutionary history and not just the initial condition.

Mathematically, the presence of speciation events reduces the number of independent single gene evolutionary events that make up the genome repertoire observed because many genomes share many moves with one another. It's clear that this mechanisms make genome more history-dependent, as evolutionary events in a single genome spread to other genomes as well.

We must remind that phylogenetic trees have themselves been called into question recently in the context of bacterial evolution. The widespread presence of gene horizontal transfer led some group to speak of phylogenetic networks instead. It is possible in principle that actually imagining independent evolution from a LUCA fits bacterial evolution fairly well. At any rate, it is necessary to understand what observable would change under the effect of clear-cut speciation and standard phylogenetic trees.

It is not actually particularly difficult to adapt the markovian framework to the presence of different clades. To generate an ensemble of genome, let us consider the following steps:

1. start from a LUCA represented by $\mathbf{n_0}$
2. generate two genomes independently using one of the models described in the previous subsection using $\mathbf{n_0}$ as initial condition; stop generating such genomes after a number of steps $m_{sp}$ extracted itself from a distribution that can be estimated from considering total domain-count changes of closely-related pair of currently-observed genomes genomes (see Nimwegen and Molina, 2008);

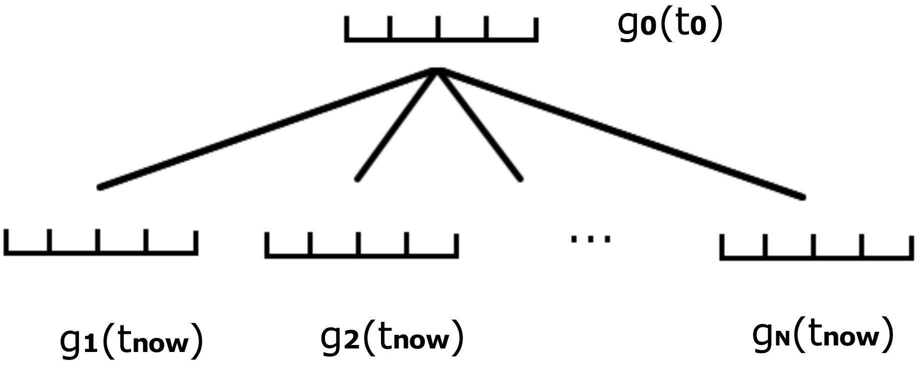

**Figure 2.5: Genomes evolving independently starting from the last common ancestor.** The class of models outlined in this chapter assume that each genome is the result of a unique evolutionary trajectory independent of the trajectories taken by the other genomes. In reality, closely-related organism share much of their evolutionary history and therefore, their trajectories are tightly related to one another.

3. for each of the two genome, generates another pair of genomes independently using their **n** as initial condition with random $m_{sp}$
4. iterate the procedure; with $k$ iterations, we will have $2^k$ genomes at the end of the evolutionary process; to have about 1000 genomes, as we observe empirically, we need $k = 10$

We can use identical family parameters for the two types of evolution, so that broad changes in the patterns observed should be an effect of the different type of evolution only.

### 2.5.1 Starwars Score

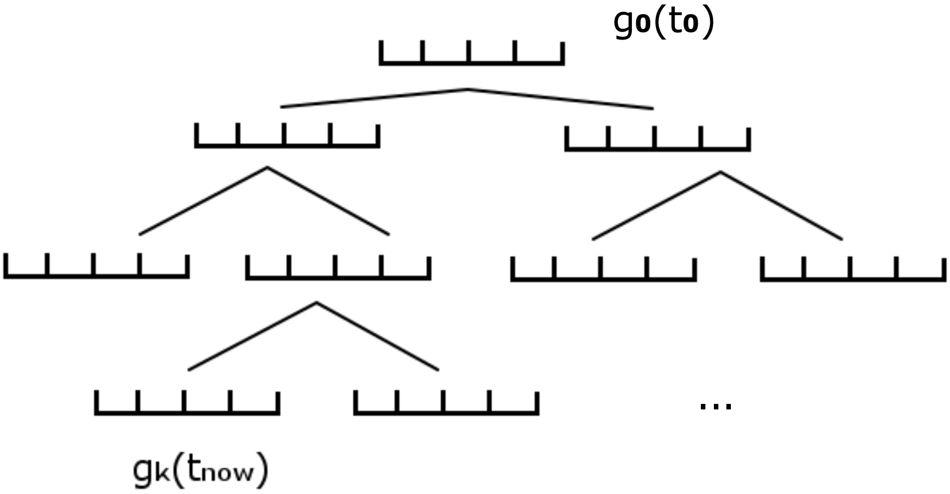

**Figure 2.6: The genomes at the end of the evolutionary process are related by intermediate common ancestors as well as the last common ancestor (LUCA) .** Each pair of genomes sharing a common ancestor has evolved from it in a manner similar to the ones described in the models of this chapter, but such models cannot technically take into account further ancestry down evolutionary history. In particular, real genomic data should present sets of genomes much more similar than it would be possible in a model of independent evolutions. Such sets are nothing other than the prokaryotic clades.

# Genome Architecture

## 3.1 Scaling

How does genome composition changes with the size of the genome? Does size increase and decrease present specific challenges of their own? Are genomes simply scaled version of one another, at least to a first approximation? How would you have to adjust genome composition if you wanted to increase its size while maintaining its functionality? These questions form the bulk of the genomic scaling theory.

The departure point of this field is the empirical observation of scaling laws in different functional categories. Such laws take the following form:

$$\langle n_f(M) \rangle = c_f M^{a_f} \tag{3.1}$$

More generally, we are interested in the whole probability distribution at fixed size:

$$P(n_1, n_2, .., n_F | M) \tag{3.2}$$

The number of gene families $F$ is itself a random variable of the size $M$ in some models. Equivalently, $F$ is a constant and we will be especially interested in the number of gene families with $n_f \neq 0$, which is always a random variable.

The structure of the distribution $P(n_1, n_2, .., n_F | M)$ reflects architectural constraints given by the genome size, itself the consequences of physical and biological principles. It is related to the fitness of a genome composition state at *fixed* size. Whether or not such composition state is actually common over the course of genome evolution will depend on the particular evolutionary pressures on size as well. Genome evolution is informed by $P(n_1, n_2, .., n_F | M)$, but it is a more complicated problem in general.

### 3.1.1 Stochastic Growth Models

Stochastic growth models sample a genome $\mathbf{n}(\mathbf{M})$ in $M$ steps. At each step, a gene is added to the genome. The gene belongs to family $f$ with probability $P(f, M)$. Such probabilities are effectively transition rates in a Markov chain. Contrary to evolutionary models, we need to consider only the possibility of gene gain:

$$(n_1, .., n_f, .., n_F) \rightarrow (n_1, .., n_f + 1, .., n_F) \tag{3.3}$$

i.e., $\mathbf{n} \rightarrow \mathbf{n_f^+}$.

In the Markov chain formalism $\pi(\mathbf{n_f^+}|\mathbf{n}) = P(f, M)$. Specifying the model is equivalent to write down such rates. It is possible to write down a master equation, but it's not very useful as the dynamics is highly non-ergodic. Instead the main analytical results come from mean field theory.

#### 3.1.1.1 A simple observation on the limits of scaling

$$M = \sum_f n_f(M) \tag{3.4}$$

$$M = \sum_f \langle n_f(M) \rangle \tag{3.5}$$

Let us suppose a scaling law for each $\langle n_f(M) \rangle$:

$$M = \sum_f c_f M^{a_f} \tag{3.6}$$

It is clear that the presence of superlinearities $a_f > 1$ would imply an upper limit on genome size as one gene family takes up the whole genome eventually. Conversely if a model is well-defined at any size, superlinear laws can hold only transiently and eventually they turn either into linear or sublinear laws. Whether or not an upper size limit actually exist in bacteria is a debated topic. Even if such limit does exist, there's no guarantee that evolution would have reached it. If it doesn't exist though, it seems safe to say that we would observe transition in the scaling behavior after a certain size.

### 3.1.2 Effective Independent Model A

We define the model in the following way:

$$P(f) = \frac{\epsilon_f n_f + p_f}{\epsilon \cdot \mathbf{n} + 1} \tag{3.7}$$

Biologically, what matters is $P(f) \propto \epsilon_f n_f + p_f$, the rest is a normalization factor. While mathematically gene families interact with one another through the normalization factor, we call this model independent in virtue of the numerator depending only on the family in consideration. Obviously, gene families are interacting with one another in reality and this is meant to be an effective model of such interactions.

The term $\epsilon_f n_f$ models the expansion mechanism dependent on the current abundance: if $\epsilon_f > 0$, it says that the more abundant the family, the more likely it is to be selected at the next step. If if $\epsilon_f < 0$, the opposite is true. The abundance-independent term $p_f$ models any abundance-independent expansion mechanism. Finally, we must have $\sum_f p_f = 1$ for $P(f)$ to be a probability.

The exact biological mechanism behind these two terms are less obvious than it might first appear. It is natural for instance to think of the term $\epsilon_f n_f$ as a model of gene duplication, while $p_f$ is a term coming from the the dynamics of gene horizontal transfer. There is some evidence in the literature, though, that the horizontal transfer rates of gene families might have a dependence on the abundance. If anything then, these models shed light on the precise phenomenology that must be obeyed by any mechanism that reproduce the empirical data.

It is clear conceptually that $\epsilon_f$ and $p_f$ are related to the fitness advantage given by a family. In fact it is tempting to consider $p_f$ proportional to the abundance of the gene family in some other space, perhaps the abstract space of all possible families or the abundance in the metagenome. Then, admittedly making a certain leap, we may write $p_f = \epsilon_f \tilde{n_f}$. This is probably much too simple, but the point is that $\epsilon_f$ and $p_f$ are not truly independent parameters and the scatter plot $(\epsilon_f, p_f)$ might exhibit some structure.

We write in mean field:

$$\Delta_M \langle n_f \rangle = \frac{\epsilon_f \langle n_f \rangle + p_f}{\epsilon \cdot \langle \mathbf{n} \rangle + 1} \tag{3.8}$$

For the sake of concision the average sign is dropped in the equation:

$$\Delta_M n_f = \frac{p_f + \epsilon_f n_f}{1 + \epsilon \cdot \langle \mathbf{n} \rangle} \tag{3.9}$$

#### 3.1.2.1 Infinite size limit

We first analyse the infinite size limit, ie, $M \to \infty$. Clearly families for which $\epsilon_f < 0$ will not contribute much because they will have saturated to $n_{max} = \frac{p_f}{|\epsilon_f|}$ as at this point $\partial_M n_f = 0$. Furthermore, families for which $\epsilon_f = 0$ do not contribute anyway to the right-hand side of the equation. Turning one's attention to families such that $\epsilon_f > 0$.

For simplicity let's say that all families have different $\epsilon_f$. The main idea is that in the $M \to \infty$ one particularly family wins out such that:

$$\epsilon_f n_f >> \epsilon_{f'} n_{f'} \qquad \forall f^{'} \tag{3.10}$$

The presence of a family asymptotically dominating all others is called *condensation*, as formally this is akin to Bose-Einstein condensation [**?**]. The end of the section considers the possibility that condensation does not happen and it will turn out that this is inconsistent with the assumption of different $\epsilon_f$. For now let us assume it and work out the consequences.

For the specific family that wins out

$$\Delta_M n_f \approx \frac{p_f + \epsilon_f n_f}{1 + \epsilon_f n_f} \approx \frac{\epsilon_f n_f}{\epsilon_f n_f} = 1 \tag{3.11}$$

meaning one obtains an asymptotic linear scaling.

For the other families $\epsilon_f > 0$

$$\Delta_M n_{f'} \approx \frac{p_{f'} + \epsilon_{f'} n_{f'}}{1 + \epsilon_f n_f} \approx \frac{\epsilon_{f'} n_{f'}}{\epsilon_f M} \tag{3.12}$$

Integrating the equation one gets

$$n_{f'} \propto M^{\frac{\epsilon_{f'}}{\epsilon_f}} \tag{3.13}$$

It is obvious that for consistency $\epsilon_{f'} < \epsilon_f$ otherwise one would have a superlinear scaling and the family $f$ could not win out. Therefore, the family that wins out is unsurprisingly the family with the greatest $\epsilon_f$.

Finally, let us consider what happens to families for which $\epsilon_f = 0$:

$$\Delta_M n_{f'} \approx \frac{p_{f'}}{1 + \epsilon_f n_f} \approx \frac{p_{f'}}{\epsilon_f M} \tag{3.14}$$

Integrating one obtains

$$n_{f'} \approxeq \frac{p_{f'}}{\epsilon_f} \log M \tag{3.15}$$

ie, logarithmic growth.

To conclude, for $M \to \infty$ the mean abundances of the families scale in the following way:

$$n_f \approxeq \begin{cases} M & \epsilon_f > 0 : \quad \epsilon_f = \max\limits_{f'} \epsilon_{f'} \\ M^{\frac{\epsilon_f}{\max_{f'} \epsilon_{f'}}} & \epsilon_f > 0 : \quad \epsilon_f < \max\limits_{f'} \epsilon_{f'} \\ \frac{p_f}{\max_{f'} \epsilon_{f'}} \log M & \epsilon_f = 0 \\ \frac{p_f}{|\epsilon_f|} & \epsilon_f < 0 \end{cases} \tag{3.16}$$

Note that these equations can possibly hold only if $n_f \gg \frac{1}{\epsilon_f}$, with $\epsilon_f$ being the greatest one among all the families. This conditions defines implicitly what it means quantitatively to have an infinite size limit and i is easily seen that $\frac{1}{\epsilon_f}$ is a kind of common relaxation scale effectively that governs the behavior of all families. No superlinear scaling is possible in this regime. The presence of such scaling in the data implies that if the model is applicable at all, we must rather be in a small size limit.

Before moving on to analysing the small size limit, let us consider the possibility that no single family wins out. To do this, let us consider a two-component model, both of which have $\epsilon_f > 0$.

$$\Delta_M n_1 = \frac{p_1 + \epsilon_1 n_1}{1 + \epsilon_1 n_1 + \epsilon_2 n_2} \tag{3.17}$$

$$\Delta_M n_2 = \frac{p_2 + \epsilon_2 n_2}{1 + \epsilon_1 n_1 + \epsilon_2 n_2} \tag{3.18}$$

If no component dominates the other in the limit $M \to \infty$, they must have the same asymptotic scaling and therefore $n_1 \approxeq c_1 M^\alpha$ and $n_2 \approxeq c_2 M^\alpha$. Then for instance for the first family we have:

$$\Delta_M n_1 \to \frac{\epsilon_1 c_1 M^\alpha}{\epsilon_1 c_1 M^\alpha + \epsilon_2 c_2 M^\alpha} = \frac{\epsilon_1 c_1}{\epsilon_1 c_1 + \epsilon_2 c_2} \tag{3.19}$$

which implies that $\alpha$ must be equal to one and that, using a similar equation for the second family as well

$$c_1 = \frac{\epsilon_1 c_1}{\epsilon_1 c_1 + \epsilon_2 c_2} \tag{3.20}$$

$$c_2 = \frac{\epsilon_2 c_2}{\epsilon_1 c_1 + \epsilon_2 c_2} \tag{3.21}$$

implying that $\epsilon_1 = \epsilon_2$. This implies that if $\epsilon_1 \neq \epsilon_2$, the two components cannot possible have the same asymptotic scaling and therefore one component will surely dominate the other. It is easy to see that essentially the same argument holds for more complicated situations with more than two components. Consequently, condensation must occur asymptotically for this model.

#### 3.1.2.2 Small size limit

Taking the small size limit consists of assuming a situation in which

$$1 \gg \sum_{f'}^{N_F} \epsilon_{f'} n_{f'} \tag{3.22}$$

so that the components are practically independent. The condition certainly implies

$$n_f \ll \frac{1}{|\epsilon_f|} \qquad \forall f \tag{3.23}$$

which again in turn must mean

$$\epsilon_f \ll 1 \qquad \forall f \tag{3.24}$$

These conditions are consistent with the analysis of the infinite size limit, as one has basically the opposite inequality 3.16. If $\epsilon_f >> 1$ for some family the abundance curve the asymptotic immediately and there is no small size limit. Having clarified what it means to realize the small size limit, the next paragraph solves the equation for the mean abundances in this limit. Given,

$$n_f(M+1) - n_f(M) \approx p_f + \epsilon_f n_f(M) \tag{3.25}$$

we iterate and expand:

$$n_f(M+1) \approx p_f + n_f(M)(1+\epsilon_f) = p_f + p_f(1+\epsilon_f) + n_f(M-1)(1+\epsilon_f)^2 + .. \tag{3.26}$$

Writing it compactly

$$n_f(M+1) \approx p_f \sum_{m=0}^{M} (1+\epsilon_f)^m \tag{3.27}$$

which it is integrated easily as it is a geometric series. Then one has:

$$n_f(M) \approx \begin{cases} \frac{p_f}{\epsilon_f}\left[(1+\epsilon_f)^M - 1\right] & \epsilon_f \neq 0 \\ p_f M & \epsilon_f = 0 \end{cases} \tag{3.28}$$

The first expression is perhaps more instructive when written out in the following way

$$n_f(M) \approx \frac{p_f}{\epsilon_f}\left[(1+\epsilon_f)^M - 1\right] \approx \frac{p_f}{\epsilon_f}\left[\exp(M\epsilon_f) - 1\right] \tag{3.29}$$

If instead of building genomes from scratch, we start from a common ancestor so to speak, then equation 3.25 has different initial conditions. More interestingly, it becomes more difficult to realize the small size limit as $\epsilon_f n_f$ starts much bigger than before. We write down the solution for generic initial conditions; whether or not this is consistent with the data will become clear later.

$$n_f(M) \approx \begin{cases} \frac{p_f}{\epsilon_f}\left[(1+\epsilon_f)^{M-M_0} - 1\right] + n_f(M_0)(1+\epsilon_f)^{M-M_0} & \epsilon_f \neq 0 \\ n_f(M_0) + p_f(M-M_0) & \epsilon_f = 0 \end{cases} \tag{3.30}$$

#### 3.1.2.3 Some Simple Examples

We consider some simple examples in this, emphasising the role of each element of the model taken by itself.

To start with, let us consider $a_f = a'_f \; \forall f$ and $p_f = 0 \; \forall f$ (no abundance-independent expansion mechanism). This is akin to a 'neutral' scaling theory as no family is favoured over any other and the initial abundances are all that matter. We will have

$$P(f) = \frac{n_f}{M} \tag{3.31}$$

then

$$\Delta n_f = \frac{n_f}{M} \tag{3.32}$$

which is easily solved giving

$$n_f(M) = \frac{n_f(M_0)}{M_0} M \tag{3.33}$$

Therefore, the neutral model predicts linear scaling laws. The presence of non-linear scaling laws implies specific constraints to be realized when changing the size of the genome: one cannot simply change a genome by adding genes belonging to random gene families. If we modify the model to allow for $p_f \neq 0$, little changes:

$$\Delta n_f = \frac{n_f + p_f}{M + 1} \tag{3.34}$$

Solving this equation by changing variable $n_f \to n_f + p_f$ one obtains again a linear expression in M:

$$n_f(M) = n_f(M_0) \frac{M+1}{M_0+1} + p_f \frac{M - M_0}{M_0 + 1} \tag{3.35}$$

Finally let us consider $p_f = 0 \; \forall f$ without any other constraints.

$$\Delta_M n_f = \frac{\epsilon_f n_f}{\epsilon \cdot \langle \mathbf{n} \rangle} \tag{3.36}$$

Then, let us consider two gene families $A$ and $B$. We can write

$$\frac{\Delta_M n_B}{\Delta_M n_A} = \frac{\epsilon_B n_B}{\epsilon_A n_A} \tag{3.37}$$

We can integrate $n_B(M)$ in terms of $n_A(M)$ by assuming small $\Delta_M n_A$ and realizing a continuous limit.

$$\frac{n_B(M)}{n_B(M_0)} = \left[ \frac{n_A(M)}{n_A(M_0)} \right]^{\frac{\epsilon_B}{\epsilon_A}} \tag{3.38}$$

This relationship is easily checked in the data as it shows up easily plotting $n_B(M)$ vs $n_A(M)$ in a log-log plot.

### 3.1.3 Effective Independent Model B

The model described in this section is a simple modification of the previous model, but as it turns out the mean-field results differ considerably. The rates are defined in the following manner:

$$P(f) = \frac{\frac{\epsilon_f n_f}{M} + p_f}{\frac{\epsilon \cdot \mathbf{n}}{M} + 1} \tag{3.39}$$

The main difference consist of making the abundance-dependent expansion mechanism proportional to the fraction of genes currently in the genome rather than the absolute number. Therefore, the probability of adding a gene belonging to family $f$ differs in genome with different sizes even if the absolute number of genes belonging to family $f$ is identical in the two genomes. If $p_f = 0$ for all gene families, then obviously there is no practical difference in the two mechanism, otherwise we will see that the system behaves rather differently. Within this model, it is natural to consider the fraction of the genome currently occupied by a certain gene family instead of the absolute abundance:

$$x_f(M) = \frac{n_f(M)}{M} \tag{3.40}$$

For $M_0 >> 1$, we can write:

$$\Delta \langle x_f \rangle = \frac{1}{M} \left\langle \frac{\epsilon_f x_f + p_f}{\epsilon \cdot \mathbf{x} + 1} \right\rangle \tag{3.41}$$

#### 3.1.3.1 Infinite size limit

By noting $\Delta \langle x_f \rangle < \frac{1}{M}$, it's clear than for sufficiently big size $\Delta \langle x_f \rangle = 0$. All the gene family fractions saturate to a certain value while their abundance grow simply linearly. This behavior is very different from the infinite size limit of the previous model. What is the saturating value? For very large M as in mean field $\Delta \langle x_f \rangle$ is constant, so is $P(f)$. Therefore $P(f)$ reaches a stationary distribution and therefore $P(f) \approx \frac{n_f(M)}{M} = x_f$. In this stead state then:

$$x_f = \frac{\epsilon_f x_f + p_f}{X} \tag{3.42}$$

with $X$ being the normalization factor. As $\sum_f x_f = 1$

$$x_f = \frac{p_f}{X - \epsilon_f} \tag{3.43}$$

As $\sum_f x_f = 1$, $X$ is implicitly defined in terms of $p_f$ and $\epsilon_f$ only.

#### 3.1.3.2 Small size limit

The small size limit gives the following equation in mean-field:

$$\Delta x_f = \frac{\epsilon_f x_f + p_f}{M} \tag{3.44}$$

This is easily solved in the continuous limit by use of the simple change of variable $x_f \to x_f + \frac{p_f}{\epsilon_f}$ and one obtains:

$$x_f(M) = x_f(M_0)\left(\frac{M}{M_0}\right)^{\epsilon_f} + \frac{p_f}{\epsilon_f}\left[\left(\frac{M}{M_0}\right)^{\epsilon_f} - 1\right] \tag{3.45}$$

### 3.1.4 Duplication/Innovation Model

The previous models assume a known universe of gene families as all the gene families are listed out in the definition of the model. A common alternative in the literature is to split the stochastic growth model into a probability $p_O(f)$ that a new gene belong to a family $f$ which is already present in the system and a probability $p_N$ of adding a gene that belongs to a family which is not already present in the system. A genome is built out of these two probabilities: at any step, either a gene is added that belongs to an old family or a gene belonging to a new family is added. Therefore, we must have $\sum_f p_O(f, M) + p_N = 1 \ \forall M$.

In the literature a common definition of the rates is the following:

$$p_O(f) = \frac{\epsilon_f n_f - \alpha}{\epsilon \cdot \mathbf{n} + \theta} \tag{3.46}$$

$$p_N = \frac{\alpha f + \theta}{\epsilon \cdot \mathbf{n} + \theta} \tag{3.47}$$

This is a version of the so-called chinese restaurant process (CRP), originally defined for $\epsilon_f$ constant for all $f$. While the terms in $\epsilon_f n_f$ have the same interpretation as the previous models, the parameters $\alpha$ and $\theta$ require further explanation. To this end, consider the number of families currently present $F(M)$. This is a random variable depending on the innovation dynamics. It's quite easy to prove that:

$$\Delta_M F = \langle p_N(M) \rangle \tag{3.48}$$

In principle $p_N(M)$ can depend on $n_f(M)$, $F(M)$ and $M$ explicitly as well. On the other hand, $\langle F(M) \rangle$ is easily accessed in the genomic data, as we simply count the number of different families in genomes of size M and take the average. It turns out that $\langle F(M) \rangle$ scales sublinearly with M, which implies decreasing rate of innovation for bigger genomes. Then, $p_N(M)$ cannot be constant in M and it's natural to think:

$$p_N(M) = \alpha \frac{F(M)}{M} \tag{3.49}$$

in order to obtain generic scaling laws $\langle F(M) \rangle \propto M^\alpha$. $\alpha$ is a therefore a parameter characterizing the innovation dynamics. After this, given $p_0(M) = \sum_f p_0(f, M) = 1 - p_N(M)$

$$p_O(M) = \frac{M - \alpha F(M)}{M} \tag{3.50}$$

so that it is natural to set

$$p_O(f, M) = \frac{n_f(M) - \alpha F(M)}{M} \tag{3.51}$$

Adding differential rates proportional to $n_f(M)$ is straightforward. Finally the $\theta$ parameter models a characteristic size $M$ needed for the $\epsilon_f n_f$ terms to set. Moreover, it allows $F(M)$ to exhibit logarithmic growth when $\alpha = 0$.

#### 3.1.4.1 Sketch of Mean-Field Results

Despite the different framework, this model is essentially equivalent to the previous Model A with respect to the mean-field population dynamics.

$$\Delta_M n_f = \frac{\epsilon_f n_f - \alpha}{\epsilon_\mathbf{f} \cdot \mathbf{n} + \theta} \tag{3.52}$$

We can easily absorb the $\theta$ parameter in the definition of $\epsilon_f$ and $\alpha$. Then:

$$\begin{aligned} \epsilon_f &\to \frac{\epsilon_f}{\theta} \\ p_f &\to \frac{\alpha}{\theta} \end{aligned} \tag{3.53}$$

At this point, all the results of section 3.1.2 are available. The only slight complication lies in the fact that one needs to consider for each family the 'time', i.e. the size, at which it was introduced, but this is easily taken into account by modifying the initial conditions.

### 3.1.5 Correlated Recipe Model

In reality, the growth of one family is undoubtedly coupled to the growth of other families. For instance, the addition of metabolic genes usually favours the addition of regulatory genes able to control the new metabolic pathways made available by novel enzymes. Therefore, a network of gene-gene interactions gives rise to the abundance statistics we observe in the data. In particular, according to this view scaling laws are the result of interactions between gene families.

Considering only two-body interaction, the correlated recipe model is defined through the following rates:

$$P(i) = \frac{\sum_{j=1}^{f} a_i j n_j + p_i}{\sum_{i,j=1}^{f} a_i j n_j + 1} \tag{3.54}$$

$a_i j$ encodes the gene-gene interaction network. In general, $a_i j$ is a function of $n_i$ and $n_j$.

To see how scaling laws are generated in this model and how $a_i j$ can be derived from biological and physical consideration, let us consider a model with two functional families only, a metabolic one and a regularory one, involving trasciption factors. In terms of evolutionary pressures, it's the access to novel metabolic pathways that confers a fitness advantage to organisms as they learn to metabolize a new nutrient or to synthesize new useful molecules. On the other hand, each metabolic gene is involved in multiple metabolic pathway as such genes can be combined to build various different pathways. Thank to this combinatorics, the number of metabolic genes is expected to scale sublinearly with the number of metabolic pathways. On the other hand, the number of metabolic transcription factor is expected to scale linearly with the number of metabolic pathways and therefore, the number of metabolic transcription factor should scale superlinearly with the number of metabolic genes. Ignoring the $p_i$ term, we can implement the consideration above by considering the mean-field ratio:

$$\frac{\Delta n_{TF}}{\Delta n_{MET}} = a(n_{TF}, n_{MET}) \frac{n_{TF}}{n_{MET}} \tag{3.55}$$

Let us imagine $\Delta n_{TF} = 1$, meaning a new metabolic pathway has been added. How many new metabolic genes, i.e., how many new enzymes were necessary for the development of such pathway? The answer is not trivial, but at the very least we know by the argument above that this number should decrease with the number of enzymes already present. It turns out that $\Delta n_{MET} = \frac{U}{n_{MET}}$ where $U$ is the total number of reactions/enzymes given by basic chemistry and physics. By plugging this in, we obtain a formula for $a(n_{TF}, n_{MET})$:

$$a(n_{TF}, n_{MET}) = \frac{n_{MET}^2}{U n_{TF}} \tag{3.56}$$

Then we can use this to integrate $n_{TF}$ in terms of $n_{MET}$ and obtain a quadratic scaling. This type of argument and direct simulation of the model prove that interactions between gene families can generate distinct scaling laws and gives perhaps a more satisfactory biological principles behind the scaling law. Thinking about the relationship between metabolic and regulatory gene for instance gives us the scaling exponent and vice versa once we understand the logic, we could have used the scaling laws to infer certain relation between the metabolic and regulatory networks of bacteria. In the same spirit one could think about for instance metabolic and sensing families are related and so on.

### 3.1.6 Coarse-Graining at the Level of Gene Family Definition

## 3.2 Family Size Distribution

## 3.3 Occurrence Distribution

## 3.4 Interactions

# Appendix

## A.1 Mean of the function and function of the mean

In this section we review how to approximate the mean of a function with the function of the mean and how to quantify the error of the approximation.

Let us consider $f(M)$ and write its mean:

$$\langle f \rangle = \int P(M) f(M) dM \tag{1}$$

We taylor-expand $f(M)$ around the mean $\langle M \rangle$:

$$f(M) = f(\langle M \rangle) + (M - \langle M \rangle) f'(\langle M \rangle) + \frac{1}{2}(M - \langle M \rangle)^2 f''(\langle M \rangle) + .. \tag{2}$$

It's immediate to obtain:

$$\langle f \rangle = f(\langle M \rangle) + \frac{\sigma_M^2}{2} f''(\langle M \rangle) + \frac{\sigma_M^3 \gamma_M}{3!} f'''(\langle M \rangle) + \frac{\sigma_M^4 K_M}{4!} f''''(\langle M \rangle) + .. \tag{3}$$

with $\sigma_M$, $\gamma_M$, $K_M$ being the standard deviation, the skewness and the kurtosis.

It is apparent that the goodness of the approximation $\langle f \rangle \approxeq f(\langle M \rangle)$ depends both on the shape of the underlying distribution $P(M)$ and the specific form of the function $f(M)$. In particular if $f(M)$ is close to being linear, the approximation is rather good.

Let us consider a function that is as far from a linear function as possible, i.e., an exponential. If $f(M) = \exp(M/\tau)$, then

$$\langle f \rangle = f(\langle M \rangle) \left[ 1 + \frac{\sigma_M^2}{2\tau^2} + \frac{\sigma_M^3 \gamma_M}{3!\tau^3} + \frac{\sigma_M^4 K_M}{4!\tau^4} + .. \right] \tag{4}$$

In our context, it's tempting to consider the observable $M$ distributed according to a gaussian distribution. In this case, the skewness is 0, the kurtosis is 3 and all that matters effectively is the ratio $\frac{\sigma_M}{\tau}$. The latter statement applies whenever the skewness and the kurtosis are of order $o(1)$. On the other hand, there is some evidence that the distribution of genome size is actually bimodal. Bimodality complicates things: for instance, in a simple symmetric gaussian mixture skewness obeys $\sigma^3\gamma = \frac{3}{2}\Delta\mu\frac{\sigma_2^2-\sigma_1^2}{2}$ and $\sigma^2 = \frac{\sigma_1^2+\sigma_2^2}{2} + \frac{(\Delta\mu)^2}{4}$, meaning skewness picks up a contribution from the mixture. Despite this, it seems highly likely that when we consider $\gamma$ by performing the appropriate ratio of these two quantities we obtain something of order $o(1)$ at worst because $(\Delta\mu)^2$ is most likely bigger than $\Delta\mu(\sigma_2^2 - \sigma_1^2)$.

We turn to analyse the meaning of the ratio $\frac{\sigma_M}{\tau}$. We have in mind:

$$f(M) = \langle n_f(M)\rangle \tag{5}$$

therefore $\tau$ gives the scale after which the change in the abundance of family $f$ becomes exponential. $\sigma_M$ is the scale of the size fluctuations. Round $\langle M\rangle$, the sizes that matter are within an interval of length $\sigma_M$. If $\sigma_M \ll \tau$, it is possible to approximate $f(M)$ linearly along this interval to $f(\langle M\rangle) + (M - \langle M\rangle)f'(\langle M\rangle)$ and therefore approximating the mean of the function with the function of mean is allowed.

This example illustrates a generic principle: it is true that $\langle f\rangle \approxeq f(\langle M\rangle)$ if $f(M)$ is approximately linear around $\langle M\rangle$ in a region of length $\sigma_M$ or greater. This criterium is major refinement over simply stating that $f(M)$ must be linear everywhere. Possible exceptions to this rule emerge with highly skewed or fat-tailed distributions. It is useful in general to think of a $\tau$ defined as the intrinsic scale of $f(M)$ within which is possible to approximate linearly around $\langle M\rangle$.

Finally let us consider the power-law case

$$f(M) = cM^\alpha \tag{6}$$

with $\alpha \le 2$. In this case the following relation holds

$$\langle f\rangle \approxeq f(\langle M\rangle) + \frac{\sigma_M^2}{2}\frac{c\alpha(\alpha-1)}{\langle M\rangle^{2-\alpha}} \tag{7}$$

which is more instructive when written

$$\langle f\rangle \approxeq f(\langle M\rangle)\left[1 + \left(\frac{\sigma_M}{\langle M\rangle}\right)^2 \alpha(\alpha-1)\right] \tag{8}$$

The sign of the correction depends on whether the power-law is sublinear or superlinear. Superlinear laws require a bigger correction. To get a feel for the numbers, consider a quadratic scaling and $\frac{\sigma_M}{\langle M\rangle} = 0.1$, then $\langle f\rangle \approxeq f(\langle M\rangle)$ up to a 2% underestimation, a very good approximation. If $\frac{\sigma_M}{\langle M\rangle} = 1$, $f(\langle M\rangle) = \frac{1}{3}\langle f\rangle$, which is instead a sever underestimation, about 66%. The main lesson to draw out of this numerical analysis is that the spread of the underlying distribution matters more than the particular exponent in the case of power-laws.

# Bibliography

[1] Van Nimwegen E, Scaling laws in the functional content of genomes *Trends Genet.* 19 (2003)

[2] Molina N *et al.*, Scaling laws in functional genome content across prokaryotic clades and lifestyles *Trends Genet.* 25(6) (2009)

[3] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV, Birth and death of protein domains: A simple model of evolution explains power law behavior, *BMC Evol Biol* 2 (2002)

[4] Cosentino Lagomarsino M *et al.*, Universal features in the genome-level evolution of protein domains *Genome Biol.* 10(1) R12 (2009)

[5] Maslov S *et al.*, Toolbox model of evolution of prokaryotic metabolic networks and their regulation, *PNAS* 106 24 (2009)

[6] Koonin EV, Wolf YI, Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World, *Nucleic Acids Research* 36 (2008)

[7] Grilli J *et al.*, Joint scaling laws in functional and evolutionary categories in prokaryotic genomes, *Nucleic Acids Research* 40 2 (2012)

[8] Grilli J *et al.*, Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers, *Phys. Rev. Lett.* 42 11 (2014)

[9] Sela I, Wolf YI, Koonin EV, Theory of prokaryotic genome evolution, *PNAS* 113 41 (2016)

[10] Mazzolini A et al., Statistics of Shared Components in Complex Component Systems, *Phys. Rev. X* 8 2 (2018)