# The quasispecies distribution

Raphaël Cerf and Joseba Dalmau
DMA, École Normale Supérieure

June 2, 2016

**Abstract**

The quasispecies model was introduced in 1971 by Manfred Eigen to discuss the first stages of life on Earth. It provides an appealing mathematical framework to study the evolution of populations in biology, for instance viruses. We present briefly the model and we focus on its stationary solutions. These formulae have a surprisingly rich combinatorial structure, involving for instance the Eulerian and Stirling numbers, as well as the up–down coefficients of permutations.

## 1 Introduction

The very concept of quasispecies is actively debated in theoretical biology. Loosely speaking, a quasispecies is a group of individuals which are closely related to each other. At the genetic level, it is a model for a cloud of mutants around a well fitted genotype, called the wild type or the master sequence. Some biologists argue that natural evolution operates on quasispecies rather than on single individuals. Ideas coming from the quasispecies theory have been successfully applied to model populations of viruses. Viruses have simple genomes which can be analyzed with modern sequencing techniques. Moreover they mutate very fast, thereby giving rise to complex quasispecies. Some medical strategies to prevent the development of viruses, like the HIV virus, are based on the quasispecies model (see [3] for a recent review). It is therefore crucial to improve our mathematical understanding of the quasispecies model, in order to derive quantitative results which can be confronted with experimental data. In this text, we shall present briefly the quasispecies model of Eigen and we shall study its stationary solutions. In doing so, we will embark on an enriching journey around a wealth of mathematical tools: Perron Frobenius theory, the polylogarithm or Jonquière's function, Eulerian and Stirling numbers, the up–down coefficients of permutations, the Poisson random walk.

## 2 The quasispecies model

Manfred Eigen introduced the quasispecies model in his celebrated article from 1971 about the first stages of life on Earth [4]. Most presumably, the first living creatures were long macromolecules. Eigen suggested that, at the macroscopic level, their evolution could be adequately described by a collection of chemical reactions. The main forces driving this evolution are selection and mutation. Accordingly, the chemical reactions model the replication or the degradation of each type of macromolecule. Moreover the replication process is subject to errors caused by mutations. Each type of macromolecule is classified according to its genotype. We denote by $E$ the set of the possible genotypes. The speed of reproduction of a macromolecule is a function of its genotype and it is given by a fitness function $f : E \to \mathbb{R}^+$. Finally, the probability that a macromolecule with genotype $u$ mutates into a macromolecule with genotype $v$ is denoted by $M(u, v)$. The concentration $x(v)$ of the genotype $v \in E$ evolves according to the differential equation

$$\frac{d}{dt} x_t(v) \;=\; \sum_{u\in E} x_t(u) f(u) M(u, v) - x_t(v) \sum_{u\in E} x_t(u) f(u)\,.$$

The first term accounts for the production of individuals having genotype $v$, production due to erroneous replication of other genotypes as well as faithful replication of itself. The negative term accounts for the loss of individuals having genotype $v$, and keeps the total concentration of individuals constant.

## 3 Stationary solutions

We shall focus on the stationary solutions of Eigen's system, that is, the solutions of the system

$$\forall u \in E \qquad x(u) \sum_{v\in E} x(v) f(v) \;=\; \sum_{v\in E} x(v) f(v) M(v, u) \qquad (\mathcal{S})$$

subject to the constraint

$$\forall u \in E \qquad x(u) \geq 0\,, \qquad \sum_{u\in E} x(u) \;=\; 1\,. \qquad (\mathcal{C})$$

Suppose that $(x(u))_{u\in E}$ is a solution to $(\mathcal{S})$ which satisfies $(\mathcal{C})$. Let $\lambda$ be the mean fitness, given by $\lambda = \sum_{v\in E} x(v) f(v)$ and let us set $y(v) = \sqrt{f(v)}x(v)$ for $v \in E$. These new variables satisfy

$$\forall u \in E \qquad \lambda y(u) \;=\; \sum_{v\in E} y(v) \sqrt{f(v)} M(v, u) \sqrt{f(u)}\,. \qquad (\mathcal{S}')$$

Therefore $(y(u))_{u\in E}$ is an eigenvector of the matrix $\sqrt{f(v)}M(v,u)\sqrt{f(u)}$. The question of the existence and uniqueness of the stationary solutions will be settled with the help of a result from linear algebra and the following hypothesis.

**Hypothesis** $(\mathcal{H})$**.** We suppose that the genotype space $E$ is finite, that the fitness function $f$ is positive, that the mutation matrix $M$ is symmetric and that all its entries are positive.

Suppose that hypothesis $(\mathcal{H})$ holds. We can apply proposition A.1 of the appendix to the matrix

$$A(u,v) \,=\, \sqrt{f(v)}M(v,u)\sqrt{f(u)}\,.$$

If $(y(u))_{u\in E}$ is a solution to $(\mathcal{S}')$ with non negative entries, then $\lambda$ has to be the largest eigenvalue of $A$ and $(y(u))_{u\in E}$ is an eigenvector associated to $\lambda$. Since the corresponding eigenspace has dimension one, there is a unique choice satisfying the constraint $(\mathcal{C})$. Therefore, under hypothesis $(\mathcal{H})$, the system $(\mathcal{S})$ admits a unique solution satisfying the constraint $(\mathcal{C})$. In fact, this result still holds if we relax the hypothesis that the mutation matrix $M$ is symmetric. We would then make appeal to the Perron–Frobenius theorem [7] to get the conclusion.

## 4 Genotypes and mutations

Ideally, we would like to have explicit formulae for $\lambda$ and $x$ in terms of $f$ and $M$. There is little hope of obtaining such explicit formulae in the general case. Therefore, we focus on a particular choice of the set of genotypes $E$ and of the mutation matrix $M$. Both for practical and historical reasons, we make the same choice as Eigen did.

**Genotypes.** We consider the different genotypes to be sequences of length $\ell \geq 1$ over the alphabet $\{\,0,1\,\}$. The space $E=\{\,0,1\,\}^\ell$ is often referred to as the $\ell$–dimensional hypercube. The hypercube is endowed with a natural distance, called the Hamming distance, which counts the number of different digits between two different sequences:

$$\forall\, u,v\in\{\,0,1\,\}^\ell \qquad d(u,v) \,=\, \text{card}\,\big\{\,1\leq i\leq \ell : u(i)\neq v(i)\,\big\}\,.$$

**Mutations.** We suppose that mutations happen independently over each site of the sequence, with probability $q\in\,]0,1[$. For $u,v\in\{\,0,1\,\}^\ell$, the mutation probability $M(u,v)$ is thus given by

$$M(u,v) \,=\, q^{d(u,v)}(1-q)^{\ell-d(u,v)}\,.$$

We have not specified the fitness function yet. Let us consider first the simplest possible scenario, a constant fitness function: $f(u) = c > 0$ for all $u \in \{\,0,1\,\}^\ell$. When the fitness function is constant, there is no selection among different genotypes, and we say that the population is selectively neutral. Under the constraint $(\mathcal{C})$, since $f$ is constant,

$$\lambda \,=\, \sum_{v\in E} x(v)f(v) \,=\, c\,.$$

With our choice of the mutation scheme, the matrix $M$ is symmetric, thanks to the symmetry of the Hamming distance. The matrix $M$ is also stochastic, that is, each row of the matrix adds up to 1. It is thus doubly stochastic, that is, each column of the matrix adds up to 1 too. We conclude that, for a constant fitness function, the unique solution of $(\mathcal{S})$ satisfying the constraint $(\mathcal{C})$ is given by

$$x(u) \,=\, \frac{1}{|E|} \,=\, \frac{1}{2^\ell}\,, \qquad u \in \{\,0,1\,\}^\ell\,.$$

However, adaptive neutrality is seldom found in biological populations. We thus embark on a quest for explicit formulae involving more complex fitness functions.

## 5 Sharp peak landscape

The simplest non neutral fitness function which comes to mind is the sharp peak: there is a privileged genotype, $w^* \in \{\,0,1\,\}^\ell$, referred to as the master sequence, which has a higher fitness than the rest. Let $\sigma > 1$ and let the fitness function $f$ be given by

$$\forall\, u \in \{\,0,1\,\}^\ell \qquad f(u) \,=\, \begin{cases} \sigma & \text{if} \quad u = w^*\,, \\ 1 & \text{if} \quad u \neq w^*\,. \end{cases}$$

This is the fitness function that Eigen studied in detail in his article [4]. One of the main advantages of working with the sharp peak is that we can break the space of genotypes into Hamming classes. For $k \in \{\,0,\dots,\ell\,\}$, the Hamming class $k$, denoted by $\mathcal{H}_k$, is the subset of $\{\,0,1\,\}^\ell$ containing all the genotypes that are at Hamming distance $k$ from the master sequence. Let us define the function $f_H : \{\,0,\dots,\ell\,\} \to \mathbb{R}^+$ by

$$\forall\, k \in \{\,0,\dots,\ell\,\} \qquad f_H(k) \,=\, \begin{cases} \sigma & \text{if} \quad k = 0\,, \\ 1 & \text{if} \quad k > 0\,. \end{cases}$$

For each $k$, the value $f_H(k)$ is the fitness common to all the genotypes in the Hamming class $k$. As the next lemma shows, the mutation probabilities can also be lumped over Hamming classes. Let $b, c \in \{0, \dots, \ell\}$ and let $X, Y$ be independent random variables with binomial distributions $X \sim \mathrm{Bin}(b, q)$, $Y \sim \mathrm{Bin}(\ell - b, q)$ and define

$$M_H(b,c) \;=\; P\big(b - X + Y = c\big)\,.$$

**Lemma 5.1** Let $b, c \in \{0, \dots, \ell\}$. For any genotype $u$ in the Hamming class $b$, we have

$$\sum_{v \in \mathcal{H}_c} M(u,v) \;=\; M_H(b,c)\,.$$

**Proof.** The quantity $\sum_{v \in \mathcal{H}_c} M(u,v)$ is the probability of $u$ ending up in the class $c$ after mutation. We call digits in a given genotype correct or incorrect depending on whether they coincide with the master sequence or not. Since $u$ is in the Hamming class $b$, it has $b$ incorrect digits and $\ell - b$ correct ones. Each digit changes state according to a Bernoulli random variable of parameter $q$. Therefore, the law of creating correct digits from the incorrect ones is $\mathrm{Bin}(b, q)$. Likewise, the law of creating incorrect digits from the correct ones is $\mathrm{Bin}(\ell - b, q)$. Noting that these binomial laws are independent of the placement of the correct and incorrect digits (and therefore of each other), we get the desired result. □

Let $k \in \{0, \dots, \ell\}$. Adding up the equations of the system $(\mathcal{S})$ for $u \in \mathcal{H}_k$ we get

$$\sum_{u \in \mathcal{H}_k} x(u) \sum_{0 \le h \le \ell} \sum_{v \in \mathcal{H}_h} x(v) f(v) \;=\; \sum_{0 \le h \le \ell} \sum_{v \in \mathcal{H}_h} x(v) f(v) \sum_{u \in \mathcal{H}_k} M(v,u)\,.$$

We set $y(k) = \sum_{u \in \mathcal{H}_k} x(u)$. In view of the above remarks, we obtain the system

$$y(k) \sum_{0 \le h \le \ell} y(h) f_H(h) \;=\; \sum_{0 \le h \le \ell} y(h) f_H(h) M_H(h,k)\,, \quad 0 \le k \le \ell\,.$$

The number of equations has been reduced from $2^\ell$ to $\ell + 1$. Moreover the new system still has the same form as $(\mathcal{S})$, and therefore all the considerations of section 3 still hold for the new system. Under the constraint $(\mathcal{C})$, the mean fitness might be rewritten as

$$\sum_{0 \le h \le \ell} y(h) f_H(h) \;=\; (\sigma - 1) y(0) + 1\,.$$

The above system becomes then

$$y(k)\big((\sigma - 1) y(0) + 1\big) \;=\; \sum_{0 \le h \le \ell} y(h) f_H(h) M_H(h,k)\,, \quad 0 \le k \le \ell\,.$$

# 6 Long chain regime

Although this system of equations is much simpler than the initial one, explicit formulae for $y$ are still out of hand. In order to get simple and useful formulae, we consider the asymptotic regime

$$\ell \to +\infty \qquad\qquad q \to 0 \qquad\qquad \ell q \to a \in ]0, +\infty[\ .$$

This asymptotic regime, already considered by Eigen, arises naturally when modeling a population of individuals with a very long genome, in which the mean number of observed mutations per individual per generation is $a$.

**Lemma 6.1** Let $b, c \geq 0$. The mutation probability $M_H(b,c)$ satisfies

$$\lim_{\substack{\ell\to\infty,\, q\to 0\\ \ell q \to a}} M_H(b,c) \;=\; \begin{cases} e^{-a}\dfrac{a^{c-b}}{(c-b)!} & \text{if} \quad c \geq b\,, \\ 0 & \text{if} \quad c < b\,. \end{cases}$$

**Proof.** Recall that if $X \sim \text{Bin}(b,q)$ and $Y \sim \text{Bin}(\ell - b, q)$ are independent random variables, then

$$M_H(b,c) \;=\; P(-X + Y = c - b)\,.$$

Since $b$ is fixed, the law $\text{Bin}(b,q)$ converges to a Dirac mass at 0, and the law $\text{Bin}(\ell - b, q)$ converges to a Poisson law of parameter $a$. The formula appearing in the lemma is precisely the probability of a Poisson random variable of parameter $a$ being equal to $c - b$. □

In view of this lemma, passing to the limit in the finite system, we obtain the infinite system of equations

$$y(k)\big((\sigma-1)y(0)+1\big) \;=\; \sum_{0\leq h\leq k} y(h) f_H(h) e^{-a}\frac{a^{k-h}}{(k-h)!}\,, \quad k \geq 0. \quad (\mathcal{S}_{sp})$$

Let's take a look at the equation for $k = 0$ first:

$$y(0)\big((\sigma-1)y(0)+1\big) \;=\; y(0)\sigma e^{-a}\,.$$

The only two solutions to this equation are

$$y(0) \;=\; 0 \qquad\qquad \text{and} \qquad\qquad y(0) \;=\; \frac{\sigma e^{-a}-1}{\sigma-1}\,.$$

On one hand, if $y(0) = 0$, it can be seen by induction that $y$ is identically 0, so this solution does not satisfy the constraint $(\mathcal{C})$. On the other hand, the second solution for $y(0)$ is positive if and only if $\sigma e^{-a} > 1$. Let us suppose that $\sigma e^{-a} > 1$, for we can only expect to find a solution satisfying the constraint $(\mathcal{C})$ in this case, and let us solve the recurrence relation defined by $(\mathcal{S}_{sp})$, with initial condition $y(0) = (\sigma e^{-a} - 1)/(\sigma - 1)$. Replacing $y(0)$ on the left hand side of $(\mathcal{S}_{sp})$ and dividing by $e^{-a}$ on both sides, the recurrence relation becomes

$$y(k)\sigma \,=\, y(0)\sigma\frac{a^k}{k!} + \sum_{1\leq h\leq k} y(h)\frac{a^{k-h}}{(k-h)!}\,, \qquad k \geq 1\,.$$

# 7 The distribution of the quasispecies

We choose to solve the recurrence relation by the method of generating functions (a beautiful account of this method can be found in chapter 7 of [5]). Set

$$g(X) \,=\, \sum_{k\geq 0} y(k)X^k\,.$$

Using the recurrence relation, we have

$$\begin{aligned} g(X)e^{aX} \,&=\, \sum_{k\geq 0}\sum_{h=0}^{k} y(h)\frac{a^{k-h}}{(k-h)!}X^k \\ &= \sum_{k\geq 0}\Big(y(k)\sigma - y(0)(\sigma-1)\frac{a^k}{k!}\Big)X^k \,=\, \sigma g(X) - y(0)(\sigma-1)e^{aX}\,. \end{aligned}$$

Replacing $y(0)$ by its value, we get

$$\begin{aligned} g(X) \,&=\, (\sigma e^{-a}-1)\frac{e^{aX}}{\sigma - e^{aX}} \,=\, (\sigma e^{-a}-1)\sum_{h\geq 1}\Big(\frac{e^{aX}}{\sigma}\Big)^h \\ &= (\sigma e^{-a}-1)\sum_{h\geq 1}\frac{1}{\sigma^h}\sum_{k\geq 0}\frac{(ah)^k}{k!}X^k \,=\, (\sigma e^{-a}-1)\sum_{k\geq 0}\frac{a^k}{k!}\sum_{h\geq 1}\frac{h^k}{\sigma^h}X^k\,. \end{aligned}$$

We deduce from here that

$$\forall\, k \geq 0 \qquad y(k) \,=\, (\sigma e^{-a}-1)\frac{a^k}{k!}\sum_{h\geq 1}\frac{h^k}{\sigma^h}\,.$$

Eigen described the quasispecies as a population of individuals having a positive concentration of the master sequence along with a cloud of mutants. We now have an explicit formula for the concentrations of the master sequence and the different mutant classes in Eigen's original quasispecies model.

**Definition 7.1** Let $\sigma, a$ be such that $\sigma e^{-a} > 1$. We say that a random variable $X$ has the distribution of the quasispecies of parameters $\sigma$ and $a$, and we write $X \sim \mathcal{Q}(\sigma, a)$, if

$$\forall k \geq 0 \qquad P(X = k) \;=\; (\sigma e^{-a} - 1)\frac{a^k}{k!}\sum_{h\geq 1}\frac{h^k}{\sigma^h}\,.$$

The above formula is a genuine probability distribution, indeed all these numbers add up to one, as can be seen by replacing $X$ by 1 in the equality

$$g(X) \;=\; (\sigma e^{-a} - 1)\frac{e^{aX}}{\sigma - e^{aX}}\,.$$

The quasispecies distribution $\mathcal{Q}(\sigma, a)$ can be expressed in terms of the polylogarithm or Jonquière's function. Let $s, z \in \mathbb{C}$, with $|z| < 1$. The polylogarithm of order $s$ and argument $z$ is defined by

$$Li_s(z) \;=\; \sum_{h\geq 1}\frac{z^h}{h^s}\,.$$

In view of this definition,

$$\forall\, k \geq 0 \qquad y(k) \;=\; (\sigma e^{-a} - 1)\frac{a^k}{k!}Li_{-k}\Big(\frac{1}{\sigma}\Big)\,.$$

Thanks to this formula, we can easily draw the quasispecies distribution and study its dependance on the parameters $\sigma, a$. In the quasispecies literature, this was previously done by integrating numerically the differential system (see for instance [3], figure 3, p.9).

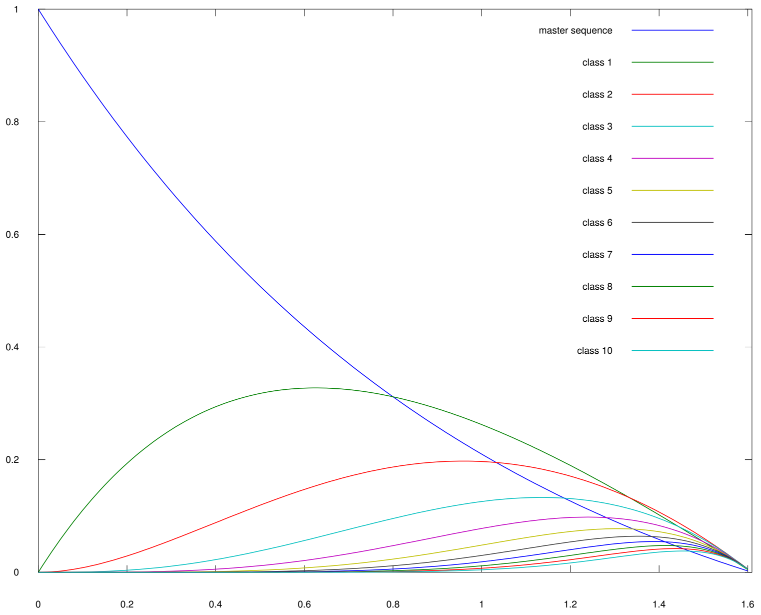
# 8 Eulerian numbers

We look next for an expression of $y(k)$ involving just a finite number of terms, instead of a series. Let $s = 1/\sigma$ and consider the well known identity

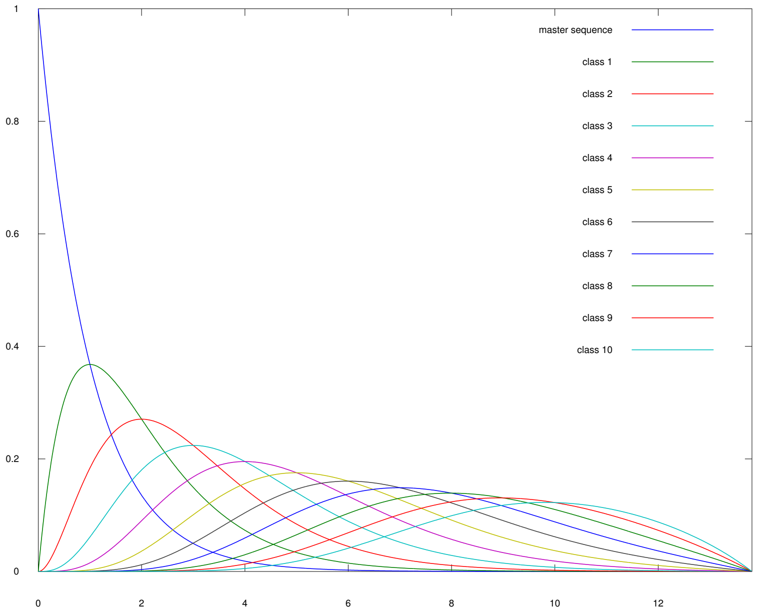
$$\sum_{h\geq 1} s^h \;=\; \frac{s}{1-s}\,.$$

We repeatedly derive and multiply by $s$ this equality, thus getting

$$\begin{aligned}\sum_{h\geq 1} hs^h \;&=\; \frac{s}{(1-s)^2}\,,\\ \sum_{h\geq 1} h^2 s^h \;&=\; \frac{s}{(1-s)^3}(1+s)\,,\end{aligned}$$



Frequency of the MS and the first 10 classes as a function of $a$ for $\sigma = 5$ .



Frequency of the MS and the first 10 classes as a function of $a$ for $\sigma = 10^6$.

$$\begin{aligned}\sum_{h\geq 1} h^3 s^h &= \frac{s}{(1-s)^4}(1+4s+s^2)\,,\\ \sum_{h\geq 1} h^4 s^h &= \frac{s}{(1-s)^5}(1+11s+11s^2+s^3)\,.\end{aligned}$$

The numbers appearing on the right hand side are the Eulerian numbers, and the polynomials are the Eulerian polynomials.

**Definition 8.1** For $0 \leq h < k$, the Eulerian number $\left\langle {k \atop h} \right\rangle$ is defined as the number of permutations of $\{\,1,\ldots,k\,\}$ having exactly $k$ ascents, that is, $k$ elements that are greater than the previous element in the permutation.

The Eulerian numbers satisfy the following identity:

$$\forall k \geq 1 \qquad \sum_{h\geq 1} h^k s^h \;=\; \frac{s}{(1-s)^{k+1}} \sum_{h=0}^{k-1} \left\langle {k \atop h} \right\rangle s^h\,.$$

Coming back to the variable $\sigma$, we get

$$\forall k \geq 1 \qquad \sum_{h\geq 1} \frac{h^k}{\sigma^h} \;=\; \frac{\sigma}{(\sigma-1)^{k+1}} \sum_{h=0}^{k-1} \left\langle {k \atop h} \right\rangle \sigma^{k-h-1}\,.$$

Using the classical identity $\left\langle {k \atop h} \right\rangle = \left\langle {k \atop k-1-h} \right\rangle$, and making the change of variable $h \to k-1-h$ in the previous sum, we can express the quantities $y(k)$ in terms of the Eulerian numbers:

$$\forall\, k \geq 1 \qquad y(k) \;=\; (\sigma e^{-a}-1)\frac{a^k}{k!}\frac{\sigma}{(\sigma-1)^{k+1}} \sum_{h=0}^{k-1} \left\langle {k \atop h} \right\rangle \sigma^h\,.$$

# 9 Stirling numbers

We have just seen that the concentration of class $k$ in the quasispecies distribution is a rational fraction in the variable $\sigma$, with denominator $(\sigma-1)^{k+1}$ and numerator $(\sigma e^{-a}-1)\sigma$ times the $k$–th Eulerian polynomial. Let us compute the partial fraction decomposition of this rational fraction. More precisely, we seek a sequence of real numbers $A_1,\ldots,A_k$ such that

$$\forall\, k \geq 1 \qquad y(k) \;=\; (\sigma e^{-a}-1)\frac{a^k}{k!}\frac{\sigma}{(\sigma-1)} \sum_{h=1}^{k} \frac{A_h}{(\sigma-1)^h}\,.$$

To find the values of the coefficients $A_h$, we write the Eulerian polynomial in terms of the powers of $(\sigma-1)$:

$$\sum_{h=0}^{k-1} \left\langle {k \atop h} \right\rangle \sigma^h \;=\; \sum_{h=0}^{k-1} \left\langle {k \atop h} \right\rangle \sum_{j=0}^{h} \binom{h}{j} (\sigma-1)^j$$

$$= \sum_{j=0}^{k-1} \left( \sum_{h=j}^{k-1} \left\langle {k \atop h} \right\rangle \binom{h}{j} \right) (\sigma - 1)^j \,.$$

**Definition 9.1** For $0 \leq h \leq k$, the Stirling number $\left\{ {k \atop h} \right\}$ is defined as the number of partitions of a set of cardinality $k$ into $h$ non empty subsets.

The Stirling and Eulerian numbers are linked through the classical identity

$$\sum_{h=j}^{k-1} \left\langle {k \atop h} \right\rangle \binom{h}{j} \;=\; (k-j)! \left\{ {k \atop k-j} \right\}.$$

See for instance [6], Proposition 5.83. Reporting in the expression involving the Eulerian polynomial, and reindexing the sum, we get

$$\forall\, k \geq 1 \qquad y(k) \;=\; (\sigma e^{-a} - 1) \frac{a^k}{k!} \frac{\sigma}{(\sigma - 1)} \sum_{h=1}^{k} \frac{h! \left\{ {k \atop h} \right\}}{(\sigma - 1)^h} \,.$$

# 10 Class–dependent fitness landscapes

We have obtained explicit formulae for the distribution of the quasispecies on the sharp peak landscape. To get these formulae, two ingredients have played a key role: the Hamming classes and the asymptotic regime. Yet, the strategy employed for the sharp peak landscape still makes sense for a wider class of fitness functions, namely, the fitness functions that only depend on the distance to the master sequence. This is a natural class of fitness functions, which is also considered in mathematical genetics. For instance, when the fitness function has a single maximum and it decreases fast with the distance, it is poetically called a mount Fujiyama type landscape. In this and the two following sections, we consider the analogue of system $(\mathcal{S}_{sp})$ for a general function $f : \mathbb{N} \to \mathbb{R}^+$:

$$y(k) \sum_{h \geq 0} y(h) f(h) \;=\; \sum_{0 \leq h \leq k} y(h) f(h) e^{-a} \frac{a^{k-h}}{(k-h)!} \,, \quad k \geq 0 \,. \qquad (\mathcal{S}_H)$$

We are only interested in the solutions of $\mathcal{S}_H$ that satisfy constraint $(\mathcal{C})$ and such that $y(0) > 0$. For if $y(0)$ is a solution of $(\mathcal{S}_H)$ with $y(0) = 0$, we can ignore the equation for $k = 0$, and the remaining system of equations falls into the form of $(\mathcal{S}_H)$ again. Thus, let us suppose that $y(0) > 0$. We look first at the equation for $k = 0$:

$$y(0) \sum_{h \geq 0} y(h) f(h) \;=\; y(0) f(0) e^{-a} \,.$$

Since we are assuming that $y(0)$ is positive, the mean fitness, given by $\sum_{h\geq 0} y(h)f(h)$, must be equal to $f(0)e^{-a}$. We make the change of variables $z(k) = y(k)/y(0)$, we replace the mean fitness by $f(0)e^{-a}$ in $(\mathcal{S}_H)$, and we divide both sides by $e^{-a}$, thus obtaining the recurrence relation

$$z(k)f(0) \;=\; \sum_{0\leq h\leq k} z(h)f(h)\frac{a^{k-h}}{(k-h)!}\,, \qquad k\geq 1\,,$$

with initial condition $z(0) = 1$. In order to get positive solutions, we make the following hypothesis.

**Hypothesis ($\mathcal{H}'$).** We suppose that the fitness of the Hamming class 0 is greater than the fitness of the other classes, i.e., $f(0) > f(k)$ for all $k \geq 1$.

This hypothesis is coherent with the Hamming class 0 corresponding to the master sequence, which is the fittest genotype. The method of generating functions cannot be implemented as easily as on the sharp peak landscape. However, it can be first guessed and then shown by induction that, for all $k \geq 1$,

$$z(k) \;=\; \frac{a^k}{k!}\frac{f(0)}{f(k)} \sum_{\substack{1\leq h\leq k\\ 0=i_0<\cdots<i_h=k}} \frac{k!}{(i_1-i_0)!\cdots(i_h-i_{h-1})!} \prod_{1\leq t\leq h} \frac{f(i_t)}{f(0)-f(i_t)}\,.$$

The probabilistic eye will perceive the key role of the Poisson distribution in this formula. We will discuss further this point in the last section.

# 11 Up–down coefficients

If we apply the previous formula to the sharp peak landscape, we recover the formula for the quasispecies involving the Stirling numbers. Indeed, in this case, the last product depends only on $h$ (it is equal to $(\sigma-1)^h$) and the sum of the multinomial coefficients is precisely equal to $h!\left\{{k \atop h}\right\}$. There is yet another formula for the quantities $y(k)$, which is the analogue of the formula involving the Eulerian numbers in the case of the sharp peak landscape. In order to present this formula, we introduce the up–down numbers or up–down coefficients. Let $n \geq 2$, and let $\sigma = (\sigma(1), \dots, \sigma(n))$ be a permutation of $1, \dots, n$. The ascents and descents of $\sigma$ are codified by the Niven signature of $\sigma$, that is, an array $(q_1, \dots, q_{n-1}) \in \{\,-1, +1\,\}^{n-1}$ such that the product $q_i(\sigma(i+1) - \sigma(i))$ is positive for all $i$. The up–down numbers, which we define next, count the number of permutations sharing the same pattern of ascents and descents.

**Definition 11.1** Let $n \geq 2$ and let $I$ be a subset of $\{\,1, \dots, n-1\,\}$. The up–down coefficient $\left\{{n \atop I}\right\}$ is defined as the number of permutations of $1, \dots, n$

having ascents in the positions $I$ and descents elsewhere. In another words, it is the number of permutations of $1,\dots,n$ having for Niven's signature

$$\forall\, i\in\{\,1,\dots,n-1\,\}\qquad q_i\;=\;\begin{cases}\;+1 & \text{if}\quad i\in I\,,\\ \;-1 & \text{if}\quad i\notin I\,.\end{cases}$$

It turns out that the quantities $z(k)$ can be expressed with the help of the up–down coefficients. For all $k\geq 1$, we have

$$z(k)\;=\;\frac{a^k}{k!}\bigg(\prod_{1\leq j\leq k}\frac{f(0)}{f(0)-f(j)}\bigg)\sum_{I\subset\{\,1,\dots,k-1\,\}}\bigg(\begin{Bmatrix}k\\ I\end{Bmatrix}\prod_{i\in I}\frac{f(i)}{f(0)}\bigg)\,.$$

In the case of the sharp peak landscape, the last product depends only on the cardinality of $I$, it is equal to $\sigma^{-|I|}$; if we sum all the terms corresponding to subsets $I$ of cardinality $h$, we obtain precisely the number of permutations of $1,\dots,k$ having $h$ ascents, which is equal to the Eulerian number $\left\langle{k\atop h}\right\rangle$.

We obtained the above formula by writing explicitly the coefficients for small values of $k$. With the help of Sloane's on–line encyclopedia of integer sequences [8], we discovered that these coefficients were the up–down coefficients. Our first proof of the formula, done in [2], relied on a difficult combinatorial identity due to Carlitz [1]. We present here a simpler more direct derivation. The strategy is to think of this formula as a rational fraction in the variables $f(1),\dots,f(k)$ and to compute its partial fraction decomposition, which turns out to be the formula given in the previous section. Thus we follow the inverse road that led us from the Eulerian numbers to the Stirling numbers when we were playing with the quasispecies on the sharp peak landscape. Let us start. We define $K=\{\,1,\dots,k-1\,\}$, we rewrite the above formula as

$$z(k)\;=\;\frac{a^k}{k!}f(0)\bigg(\prod_{1\leq j\leq k}\frac{1}{f(0)-f(j)}\bigg)\sum_{I\subset K}\bigg(\begin{Bmatrix}k\\ I\end{Bmatrix}f(0)^{k-|I|-1}\prod_{i\in I}f(i)\bigg)$$

and we expand the power $f(0)^{k-|I|-1}$ as

$$\begin{aligned}f(0)^{k-|I|-1}\;&=\;\prod_{j\in K\setminus I}\big(f(0)-f(j)+f(j)\big)\\ &=\;\sum_{J\subset K\setminus I}\bigg(\prod_{j\in J}\big(f(0)-f(j)\big)\bigg)\bigg(\prod_{j\in(K\setminus I)\setminus J}f(j)\bigg)\,.\end{aligned}$$

Reporting and simplifying the factors $(f(0)-f(j))$, we obtain

$$z(k)\;=\;\frac{a^k}{k!}f(0)\sum_{I\subset K}\sum_{J\subset K\setminus I}\bigg(\prod_{j\in K\cup\{k\}\setminus J}\frac{1}{f(0)-f(j)}\bigg)\bigg(\begin{Bmatrix}k\\ I\end{Bmatrix}\prod_{j\in K\setminus J}f(j)\bigg)\,.$$

We reindex the sum by setting $H = K \setminus J$ and we get

$$z(k) \,=\, \frac{a^k f(0)}{k! f(k)} \sum_{H \subset K} \Bigg( \prod_{j \in H \cup \{k\}} \frac{f(j)}{f(0) - f(j)} \Bigg) \Bigg( \sum_{I \subset H} \left\{ \begin{matrix} k \\ I \end{matrix} \right\} \Bigg) .$$

Let us fix $H \subset K$, say $H = \{ i_1, \dots, i_{h-1} \}$, where $1 \leq h \leq k$ and

$$i_0 = 1 \leq i_1 < \cdots < i_{h-1} < k = i_h \,,$$

and let us focus on the last sum $\sum_{I \subset H} \left\{ \begin{smallmatrix} k \\ I \end{smallmatrix} \right\}$. This sum is the number of permutations of $1, \dots, k$ whose ascents are located in the index set $H$. Let $B = \big(B_1, \dots, B_h\big)$ be an ordered partition of $\{ 1, \dots, k \}$ in $h$ subsets such that

$$\forall j \in \{ 1, \dots, h \} \qquad |B_j| = i_j - i_{j-1} \,.$$

We list the elements of each set $B_j$ in decreasing order:

$$\forall j \in \{ 1, \dots, h \} \qquad B_j \,=\, \big(b_j(1), \dots, b_j(i_j - i_{j-1})\big) \,.$$

We concatenate these lists into a single sequence:

$$b_1(1), \dots, b_1(i_1), b_2(1), \dots, b_2(i_2 - i_1), \dots, b_h(1), \dots, b_h(i_h - i_{h-1}) \,.$$

This sequence corresponds to a permutation of $1, \dots, k$. This construction defines a one to one correspondence between ordered partitions of $\{ 1, \dots, k \}$ into $h$ subsets of respective sizes $i_1, \dots, i_h - i_{h-1}$ and the set of the permutations of $1, \dots, k$ whose ascents are located in the index set $H$. The number of these partitions (called $h$–sharing in the terminology of [6], see definition 1.17 and proposition 5.5 therein) is precisely the multinomial coefficient $\frac{k!}{(i_1 - i_0)! \cdots (i_h - i_{h-1})!}$ and we conclude that

$$\sum_{I \subset H} \left\{ \begin{matrix} k \\ I \end{matrix} \right\} \,=\, \frac{k!}{(i_1 - i_0)! \cdots (i_h - i_{h-1})!} \,.$$

In fact, this combinatorial identity and the above argument are the starting point of Carlitz work [1]. The goal of Carlitz was to invert this formula, i.e., to express the up–down coefficients as sums of multinomial coefficients. Plugging this identity in the formula for $z(k)$, we are back to the formula obtained by induction in section 10.

## 12 Re–expansion

Our first formula for the distribution of the quasispecies, given in definition 7.1, was a series. Summing this series, we got a closed formula involving the Eulerian numbers. With the help of a classical combinatorial identity, we could rewrite this formula in terms of the Stirling numbers. On the class–dependent fitness landscapes, these last two formlas were generalized into two formulae, one involving multinomial coefficients, the other involving up–down coefficients. We shall finally expand these two formulae in a series in order to obtain the generalization of our first quasispecies formula. The strategy is straightforward. We expand each fraction as a geometric series. Starting from the formula involving the up–down coefficients, we obtain directly

$$z(k) \,=\, \frac{a^k}{k!}\sum_{j_1,\dots,j_k\geq 0}\bigg(\frac{f(1)}{f(0)}\bigg)^{j_1}\cdots\bigg(\frac{f(k)}{f(0)}\bigg)^{j_k}\sum_{I\subset\{\,1\leq i<k:j_i\geq 1\,\}}\begin{Bmatrix}k\\I\end{Bmatrix}.$$

Starting from the formula involving the multinomial coefficients, we obtain, after reordering the summation and setting $s_1=i_1,\dots,s_h=i_h-i_{h-1}$,

$$z(k) \,=\, \sum_{\substack{1\leq h\leq k\\ j_1,\dots,j_k\geq 1}}\;\sum_{\substack{s_1,\dots,s_h\geq 1\\ s_1+\cdots+s_h=k}}\frac{a^k}{s_1!\cdots s_h!}\bigg(\frac{f(s_1)}{f(0)}\bigg)^{j_1}\cdots\bigg(\frac{f(s_1+\cdots+s_h)}{f(0)}\bigg)^{j_h}.$$

## 13 The killed Poisson walk

The Poisson distribution appeared recurrently in our formulae for the quasispecies distribution. Let $a>0$ and let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables distributed according to the Poisson law of parameter $a$:

$$\forall n\geq 1\quad\forall k\geq 0\qquad P(X_n=k)\,=\,e^{-a}\frac{a^k}{k!}\,.$$

We consider the associated random walk on the non–negative integers, given by $S_0=0$ and

$$\forall n\geq 1\qquad S_n\,=\,X_1+\cdots+X_n\,.$$

Our final goal is to obtain a probabilistic representation of the quasispecies distribution in terms of the Poisson random walk killed at a random time. More precisely, we aim at constructing an integer–valued random variable $\tau$ such that the concentrations $\big(y(k)\big)_{k\geq 0}$ of the Hamming classes in the quasispecies are equal to the mean empirical distribution of the Poisson random walk between times 1 and $\tau$, that is,

$$\forall k\geq 0\qquad y(k)\,=\,\frac{1}{E(\tau)}E\bigg(\sum_{n=1}^{\tau}1_{\{S_n=k\}}\bigg)\,.\qquad(\diamondsuit)$$

We start this program on the sharp peak landscape.

**Proposition 13.1** Let $\sigma, a$ be such that $\sigma e^{-a} > 1$. Let $\tau$ be a random variable, which is independent of the Poisson random walk, with geometric distribution of parameter $1-(\sigma e^{-a})^{-1}$. With this choice of $\tau$, the probabilistic representation ($\diamondsuit$) holds for the quasispecies distribution $\mathcal{Q}(\sigma, a)$.

We recall that the geometric distribution of parameter $1-(\sigma e^{-a})^{-1}$ is

$$\forall n \geq 1 \qquad P(\tau \geq n) \,=\, \Big(\frac{1}{\sigma e^{-a}}\Big)^{n-1}.$$

**Proof.** We compute the expectation by decomposing the sum according to the value of $\tau$:

$$\begin{aligned} E\Big(\sum_{n=1}^{\tau} 1_{\{S_n=k\}}\Big) \,&=\, \sum_{t=1}^{\infty} E\Big(\sum_{n=1}^{t} 1_{\{S_n=k,\,\tau=t\}}\Big) \\ &= \sum_{t=1}^{\infty}\sum_{n=1}^{t} P\big(S_n=k,\,\tau=t\big) \,=\, \sum_{n=1}^{\infty}\sum_{t=n}^{\infty} P\big(S_n=k,\,\tau=t\big)\,. \end{aligned}$$

Now, the variables $S_n$ and $\tau$ are independent. The distribution of $\tau$ is geometric, the distribution of $S_n$ is Poisson of parameter $na$ (it is a sum of $n$ independent Poisson distributions of parameter $a$). Thus the previous sums become

$$\begin{aligned} \sum_{n=1}^{\infty}\sum_{t=n}^{\infty} P\big(S_n=k\big)P\big(\tau=t\big) \,&=\, \sum_{n=1}^{\infty} P\big(S_n=k\big)P\big(\tau\geq n\big) \\ &= \sum_{n=1}^{\infty} e^{-na}\frac{(an)^k}{k!}\Big(\frac{1}{\sigma e^{-a}}\Big)^{n-1} \,=\, \sigma e^{-a}\frac{a^k}{k!}\sum_{n=1}^{\infty}\frac{n^k}{\sigma^n}\,. \end{aligned}$$

Since $E(\tau) = \sigma e^{-a}/(\sigma e^{-a}-1)$, we recover the quasispecies distribution on the sharp peak landscape. □

The previous construction can be extended to a class–dependent fitness as follows. Suppose that at time $n$, the random walk $S_n$ is in state $i \geq 1$. We toss an independent coin of parameter $e^a f(i)/f(0)$ to decide whether the walk survives another unit of time or not. More precisely, we define, for any $i, n \geq 0$,

$$P\big(\tau \geq n+1 \,\big|\, S_n=i,\,\tau\geq n\big) \,=\, \begin{cases} 1 & \text{if} \quad i=0\,, \\ e^a\,\dfrac{f(i)}{f(0)} & \text{if} \quad i\geq 1\,. \end{cases}$$

This defines a random time $\tau$ whose distribution is a predictable function of the trajectory of the Poisson walk $(S_n)_{n\in\mathbb{N}}$, i.e., the event $\tau=n+1$ depends on $n$ independent coins whose parameters are deterministic functions of the trajectory $S_0, \dots, S_n$ until time $n$. Of course, the definition of $\tau$ makes sense only when $f(0) \geq e^a f(k)$ for all $k \geq 1$. With these choices, the probabilistic representation $(\diamondsuit)$ holds for the quasispecies distribution associated to $f$.

Our probabilistic construction provides the following intuitive picture for the structure of the quasispecies. The evolution of the genotype along a lineage is modelled by a Poisson random walk in the genotype space, starting from the master sequence. Because of the presence of the master sequence in the population, the lineages are bound to become extinct, after a random time which depends on their fitness history. A lineage is more robust if it visits genotypes whose fitnesses are close to the fitness of the master sequence. The time $\tau$ models the survival time of a lineage.

**Acknowledgements:** This work was completed during a visit to the mathematics department of the University of Padova. We warmly thank Carlo Mariconda and Paolo Dai Pra for their hospitality.

# A Appendix

**Proposition A.1** Let $A$ be a square matrix, which is symmetric, and whose entries are all positive. Then its eigenvalues are real, the largest eigenvalue $\lambda$ is positive, and the corresponding eigenspace has dimension one. Moreover there exists an eigenvector associated to $\lambda$ whose coordinates are all positive. Finally any eigenvector of $A$ whose coordinates are all non–negative is associated to $\lambda$.

**Proof.** Since $A$ is symmetric and real, all its eigenvalues are real. The sum of its eigenvalues is equal to its trace, which is positive, thus the largest eigenvalue of $A$ is positive. Let us call it $\lambda$. Let $y = (y(u))_{u \in E}$ be an eigenvector associated to $\lambda$:

$$\forall u \in E \qquad \lambda y(u) \,=\, \sum_{v \in E} A(u,v) y(v)\,.$$

We can assume that the Euclidean norm of $y$ is 1, i.e., $\langle y, y \rangle = 1$, where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in $\mathbb{R}^E$. Multiplying the previous equation by $y(u)$ and summing over $u \in E$, we get

$$\lambda \,=\, \sum_{u,v \in E} y(u) A(u,v) y(v) \,=\, \langle y, Ay \rangle\,.$$

Let us denote by $|y|$ the vector $(|y(u)|)_{u \in E}$. Since the entries of $A$ are positive, we deduce from the previous identity that

$$\lambda \,=\, \langle y, Ay \rangle \,\leq\, \langle |y|, A|y| \rangle \,\leq\, \sup_{z : \langle z, z \rangle = 1} \langle z, Az \rangle\,.$$

However, since $A$ is symmetric real, the last supremum is precisely equal to $\lambda$. Therefore all the previous inequalities were in fact equalities. Since all the entries of $A$ are positive, we conclude that all the entries of $y$ have the same sign. The eigenvector identity implies furthermore that no entry of $y$ is null. So far, we have prove than an eigenvector associated to $\lambda$ has all its entries positive, or all negative, and none of them is zero. Let $y, z$ be two eigenvectors associated to $\lambda$. We choose $\alpha$ to that the first coordinate of $y - \alpha z$ vanishes. Since we have $A(y - \alpha z) = \lambda(y - \alpha z)$, necessarily $y - \alpha z = 0$. Thus the eigenspace associated to $\lambda$ has dimension one. Finally, let $y$ be an eigenvector associated to $\lambda$ with positive coordinates and let $z$ be another eigenvector of $A$ with non–negative coordinates, associated to an eigenvalue $\mu$. We can find $\alpha > 0$ sufficiently small so that $z(u) \geq \alpha y(u)$ for $u \in E$. We have then, for any $n \geq 1$,

$$\langle z, A^n z\rangle \,=\, \mu^n \langle z, z\rangle \,\geq\, \langle \alpha y, A^n \alpha y\rangle \,=\, \alpha^2 \lambda^n \langle y, y\rangle\,.$$

Sending $n$ to infinity, we conclude that $\mu \geq \lambda$, therefore $\mu = \lambda$. □

# References

[1] L. Carlitz. Permutations with prescribed pattern. *Math. Nachr.*, 58:31–53, 1973.

[2] Raphaël Cerf and Joseba Dalmau. Quasispecies for class–dependent fitness landscapes. *ArXiv e-prints*, 2015.

[3] E. Domingo and P. Schuster. *Quasispecies: From Theory to Experimental Systems*. Current Topics in Microbiology and Immunology. Springer International Publishing, 2016.

[4] Manfred Eigen. Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[5] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics*. Addison-Wesley Publishing Company, Reading, MA, second edition, 1994. A foundation for computer science.

[6] Carlo Mariconda and Alberto Tonolo. *Discrete calculus –Methods for counting–*. Springer, 2016 (to appear).

[7] E. Seneta. *Nonnegative matrices and Markov chains*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1981.

[8] N. J. A. Sloane. The on-line encyclopedia of integer sequences. *Ann. Math. Inform.*, 41:219–234, 2013.