Evolution is Mathematically Determined

Daniil Demidov

19.02.2025

Abstract

We mathematically formalize the core idea of evolution—natural selection—and show that it is not only a theory of a natural sciences but rather a logically and mathematically accurate concept.

1 Introduction

A lot of mathematicians (primarily religious) state that evolution is mathematically almost certainly unlikely providing the following argument:

> The genome of an individual species can be considered as a finite sequence $g_1, ..., g_n$ for some $n \in \mathbb{N}$ bein the amount nucleotides in the DNA molecule (this is our way of saying *the number of nucleotides*) encoding all necessary information to replicate itself, whereby each g_j is an element of the set $\{A, G, T, C\}$. Thus, assuming that evolution (especially at the initial stages) is based on random changes to the genome (a.k.a. mutations), the probability of a given (or even any "livable") sequence to occur lies around 4^{-N}, where N is the length of the DNA. Considering modern scientific data, the length of the DNA molecule of an average *homo sapiens* is $3.1 \cdot 10^9$, so the probability of a human being arising through the process of evolution is $4^{-3.1 \cdot 10^9}$
> Аргумент Савватеева?

But this view has largely been criticized by professional biologists (references!) which, although, had no effect on the worldview of the mathematicians denying them. Honestly, we mathematicians, indeed, rarely listen to even plausible arguments if they are not *mathematical* and conflict with our own understanding of the reality, especially if it is backed by some formal *mathematical* argument.

So what we aim for, is to demonstrate on a simple model that the idea of evolution is indeed mathematically backed up and is not only plausible but *natural* and that it comes in one bundle with our understanding of logic and mathematics.

2 Natural Selection

The key concept which modern evolution theory is based on is *natural selection*. It is the idea that, although mutations occur randomly, the observed "result" of that process is not just a random outcome: over the generations, *good* (meaning the most suitable for the environment) mutations are most likely preserved, while bad ones are ruled out just because they less likely lead to a specimen reproducing and passing its nucleotides containing that *bad* mutation to the future generations, thus "getting rid" of it in the long run.

To me, that sounds like a strongly logical argument which can and should be mathematically formalized. That would allow us rigorously prove that, indeed, evolution

is not just a theory but a strong logical concept, backed up by mathematical tools, just like the existence of π or e, or that no quintic polynomial is generally solvable in radicals, or… You got the point.

2.1 The Simplest Model

Of course, every probabilistic analysis requires careful modelling. We will start with a simple one, potentially adding more and more details to it.

Definition 2.1.1: *(Genetic Alphabet, Nucleotides)* A finite set of symbols Γ is called *genetic alphabet*. In our case, $\Gamma = \{A, G, T, C\}$. Its elements are called **nucleotides**.

Definition 2.1.2: *(Genome)* A finite ordered collection of nucleotides $\{g_j\}$, whereby $j \in \{1, ..., N\}$, over a given alphabet Γ is called **genome**. $N \in \mathbb{N}$ is then its **length**.

We will model evolution of one given genome of length $N \in \mathbb{N}$ by looking at the sequence of genomes, such that each pair of consecutive genomes represents two consecutive generations.

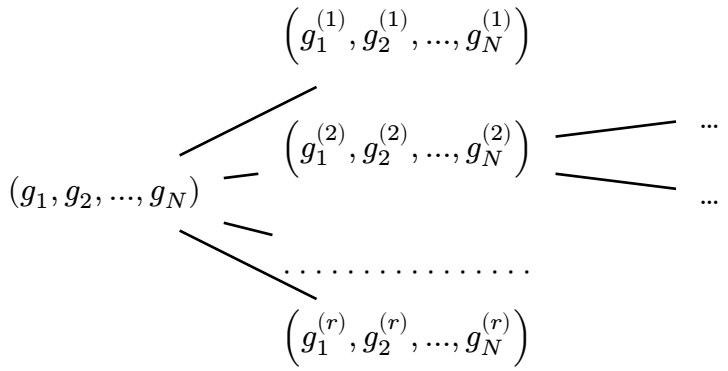
Definition 2.1.3: *(Environment)* A function $\eta : \Gamma^N \to [0, 1]$ is called **environment**. It represents the environmental factor, indicating how suitable a given genome is to it. 1 represents a perfect adaptation.

2.1.1 Assumptions

Let's outline the main principles of our modelling:

(i) The **length** of the genome is **constant**;
(ii) The **environment is constant** (this will be handled in a more complicated model later and actually shown to be not that important);
(iii) Each nucleotide has a probability $\mu \in (0, 1)$ of mutating—being replaced by one of the 3 other nucleotides;
(iv) Each genome has $r \in \mathbb{N}$ **offsprings** in the next generations, of which **the most suitable one** (with the highest value of η) survives and is considered the next element of our sequence;
(v) Each nucleotide mutates independently from other genes.

So the process goes something like this:



E.g. $\mathbb{P}(g_2^1 = g_2) = 1 - \mu$ (the probability that no mutation occurred). Say $\eta\left(\left\{g_j^{(2)}\right\}\right) = \max_{k \in \{1,...,r\}} \eta\left(\left\{g_j^{(r)}\right\}\right) \rightsquigarrow$ the genome $\left(g_1^{(2)}, g_2^{(2)}, ..., g_N^{(2)}\right)$ survives and is taken as the mutating genome of the next generation, to which the same mechanism is applied.

2.1.2 Individual Nucleotides

Considering the last assumption, let's take a look at one nucleotide's probability of being the one maximizing η.

For that sake, we assume that there exists a genome $G_{opt} \in \Gamma^N$, such that

$$\eta(G_{opt}) = \max_{G \in \Gamma^N} ! \tag{1}$$

.

We take one arbitrary nucleotide at the position $j \in \{1, ..., N\}$ and look the probability of it being the same as $(G_{opt})_j$ which ensures it being the most suitable for the given environment. We denote that probability in the k-th generation as $\pi_k \in [0, 1]$. If a mutation occures, such that $g_j = (G_{opt})_j$ in some generation, we call this mutation **useful**.

Another assumption that we need to make is that the probability of this nucleotide mutating in the next generation is some constant $\mu \in (0, 1)$.

We start our observations by looking at some genome that is considered the *first* generation. We assume it being completely random, as it is the initial point of the process of evolution. Clearly, $\pi_1 = \frac{1}{|\Gamma|} = \frac{1}{4}$.

Now we want to *inductively* define the sequence. Assume π_k is known. Then, this organism will have $r \in \mathbb{N}$ offsprings as we stated earlier. That ultimately means that our j-th nucleotide will **not** mutate in at least one of the offsprings with the probability of $1 - \mu^r$. On the other hand, if this nucleotide is not *useful* (which has the probability of $1 - \pi_k$), given that at least one mutation occurs, the probability of it being useful in the next generation is then $\frac{1}{3}$.

By definition,

$$\mathbb{P}(\text{M.}^1 \text{ is useful} \mid \text{a.l.}^2 \text{ one M. occurred}) = \frac{\mathbb{P}(\text{M. is useful} \cap \text{a.l. one M. occurred})}{\mathbb{P}(\text{a.l. one M. occurred})} \tag{2}$$

so by considering that $\mathbb{P}(\text{a.l. one M. occurred}) = 1 - (1 - \mu)^r$, we obtain

$$\mathbb{P}(\text{M. useful} \cap \text{a.l. one M. occurred}) = \frac{1}{3}(1 - (1 - \mu)^r). \tag{3}$$

Similarly, by definition

$$\begin{aligned} &\mathbb{P}(\text{N. is useful in generation } k+1) = \\ &\quad \mathbb{P}(\text{N. is useful in generation } k) \cdot \mathbb{P}(\text{no mutation occurred}) + \\ &+\mathbb{P}(\text{N. is not useful in generation } k) \cdot \mathbb{P}(\text{M. useful} \cap \text{a.l. one M. occurred}), \end{aligned} \tag{4}$$

which is equivalent to

$$\pi_{k+1} = \pi_k(1 - \mu^r) + (1 - \pi_k)\frac{1}{3}(1 - (1 - \mu)^r). \tag{5}$$

For convenience purposes, we write $a := 1 - \mu^r$ and $b := 1 - (1 - \mu)^r$. Then $a, b \in (0, 1)$, so that we have

$$\pi_{k+1} = a\pi_k + \frac{1}{3}b(1 - \pi_k). \tag{6}$$

[1] mutation

[2] at least

2.1.3 Convergence & Limit

To show that the sequence converges, we define $f(x) := x\big(a - \frac{1}{3}b\big) + \frac{1}{3}b$. Then we have $\pi_{k+1} = f(\pi_k) = f(f(...f(\pi_1)))$ for all $k \in \mathbb{N}$. Clearly,

$$f'(x) = a - \frac{1}{3}b = 1 - \mu^r - \frac{1-(1-\mu)^r}{3} \in (0,1) \tag{7}$$

for plausible values of $0 < \mu \ll 1$ and $r \geq 2$. There also exists a fixed point $\pi \in \mathbb{R}$ such that $\pi = f(\pi)$. It is not hard to see that $\pi = 1 - \frac{3\mu^r}{1-(1-\mu^r)} \in (0,1)$ and $\pi_1 = \frac{1}{4} < \pi$. Thus, π_k converges to π (fixed-point convergence theorem prove for this case? – I have a proof, but is it necessary here?).

If we consider $\mu \to 0$, then $\pi = 1 - 3\Big(\mu^{-r} - \big(\frac{1}{\mu} - 1\big)^r\Big)^{-1} \to 1$ (this limit is left as an exercise to the reader).

That alone shows that each nucleotide has a tendency of stabilizing in its most suitable state **if** the environment does not change and the conditions above are met.

2.1.4 The Expected Value of Suitable nucleotides in the Whole Genome

Let X_k be random variables indicating the amount of suitable nucleotides in the k-th generation. What is then $\mathbb{P}(X_k = m)$ for some $m \in \mathbb{N}$?

It is not hard to see that X_k is a binomially distributed random variable with $X_k \sim \text{Bin}(N, \pi_k)$. Thus, $\mathbb{E}X = N\pi_k \to N\Big(1 - 3\frac{\mu^r}{1-(1-\mu)^r}\Big) \approx N \cdot 1 = N$ as we have seen earlier. This concludes our analysis of the first model showing that even such a simple idea of natural selection is indeed mathematically justified.