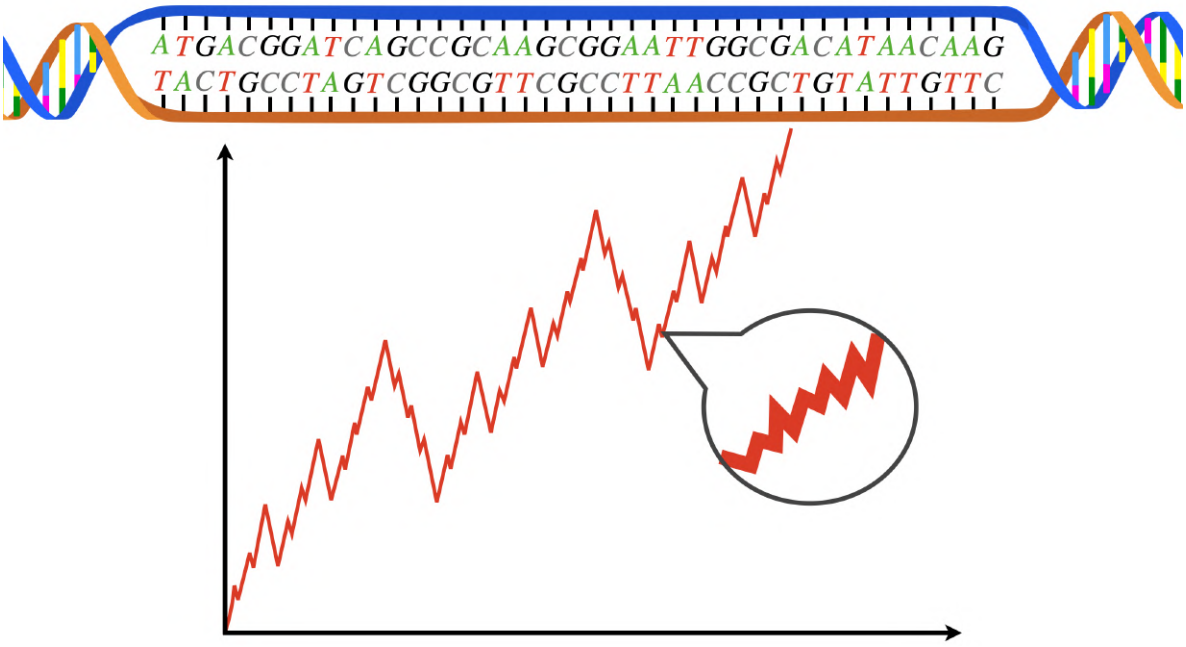
The Fractal That Lives in Your DNA

by DiBeos

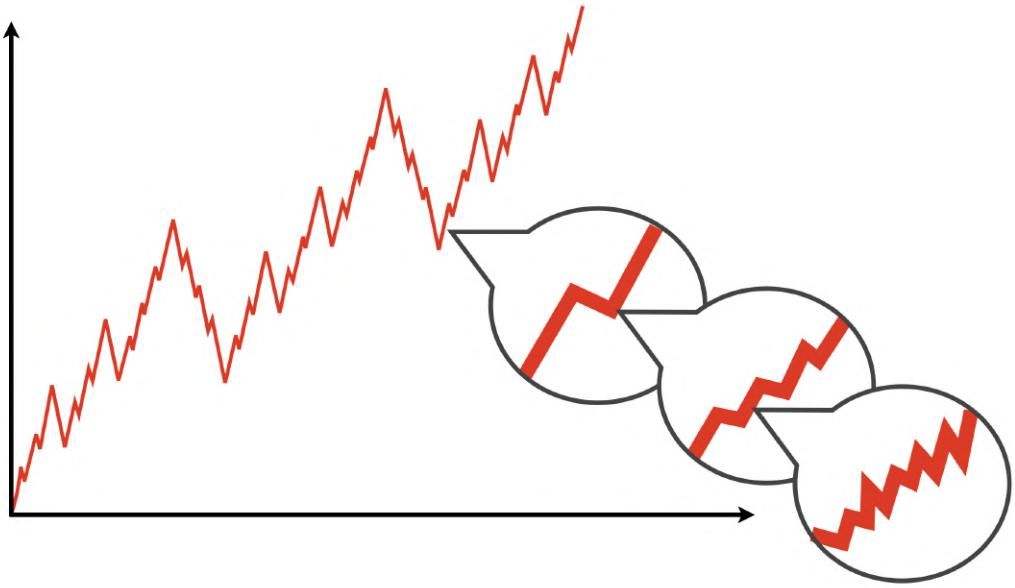


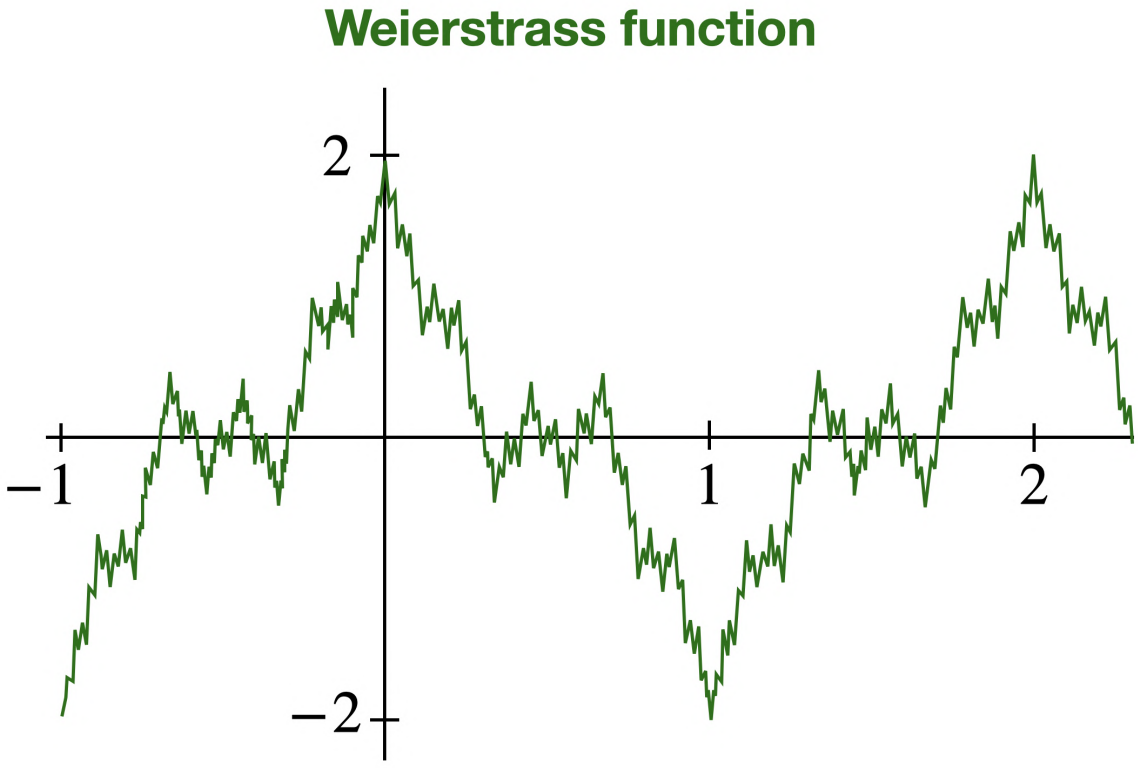
"The book of nature is written in the language of mathematics."
– Galileo Galilei

Introduction

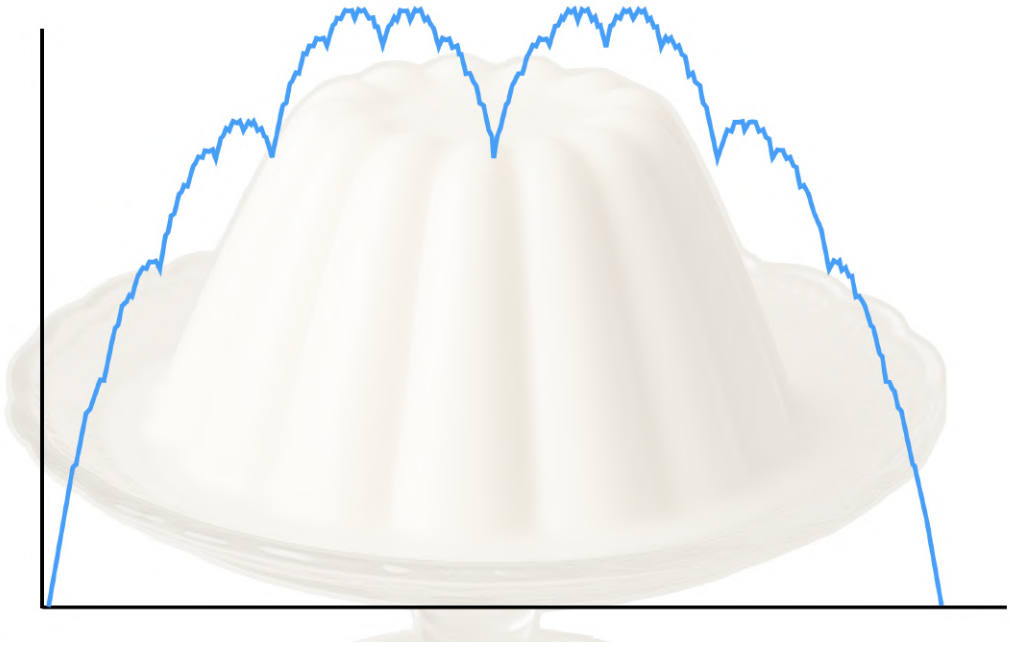
What if I told you that a simple function from number theory determines how resistant your DNA is to random mutations, and possibly even the survival of life itself?

But that's not all. This function turns out to be continuous everywhere, but differentiable nowhere. It is a fractal. One of the so-called *pathological functions* in analysis, similar to the famous *Weierstrass function*.





But it actually looks like a pudding, or more precisely a European dessert called *blancmange*. In fact, mathematicians describe this fractal as a blancmange-like curve. But more on this later...



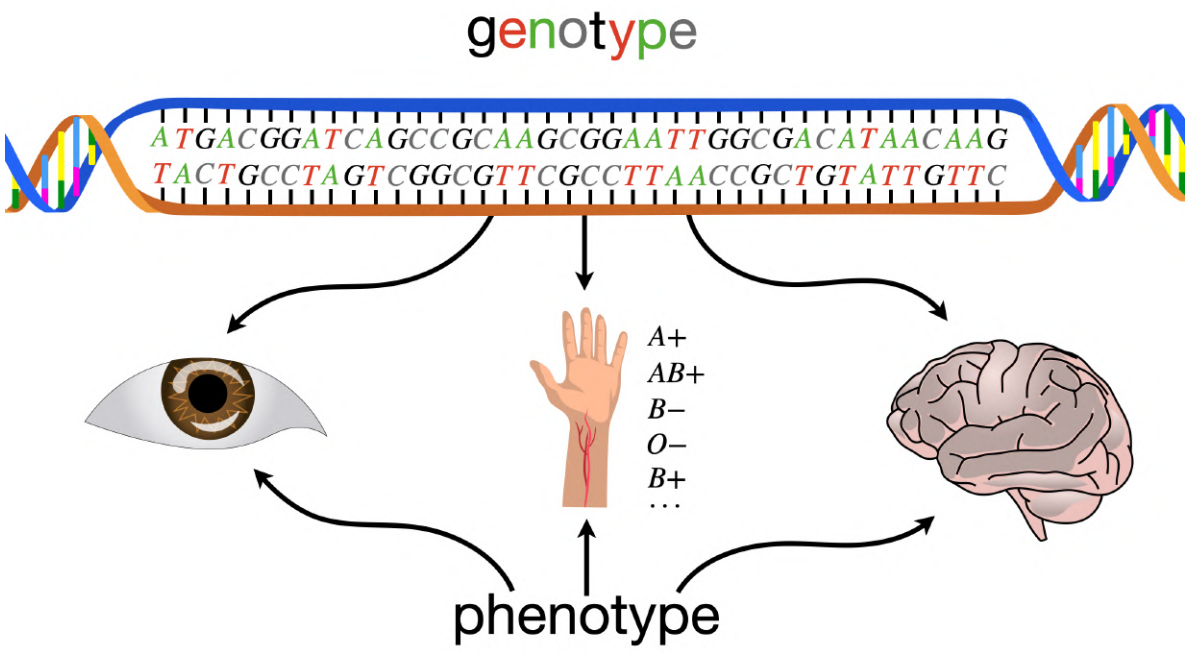
Before starting, I just want to say that we, the ***DiBeos***, are not biologists. And in this PDF we will be focusing on math, so if you find any imprecision when we talk about biology, you can check the *research paper* that this document is based on for precise biological concepts.

Let's start at the beginning:

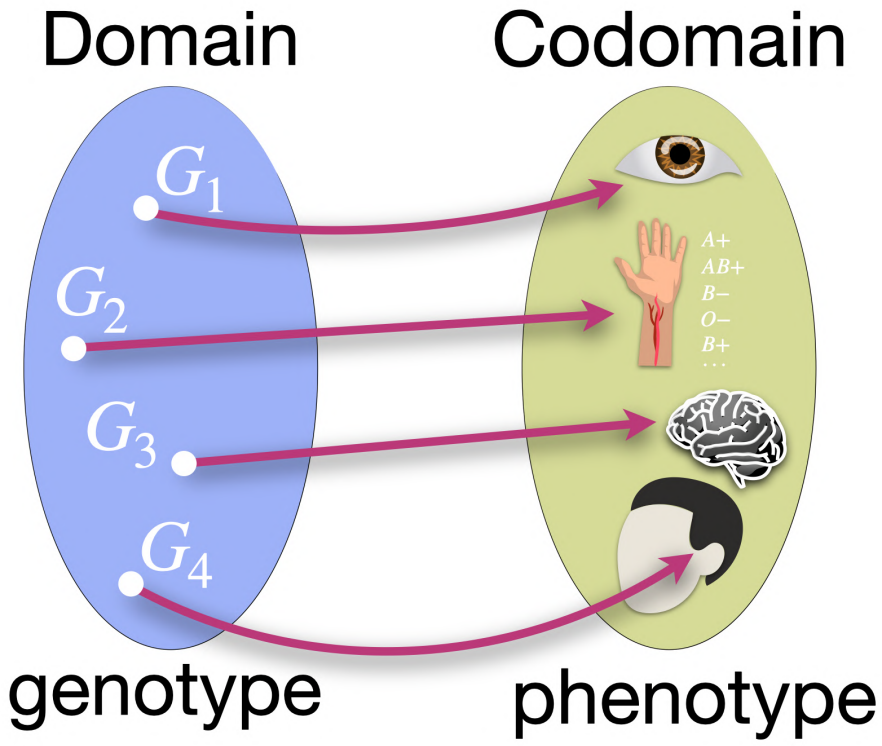
The Genotype-Phenotype Mapping

Your **genotype** (loosely speaking) is your DNA sequence.

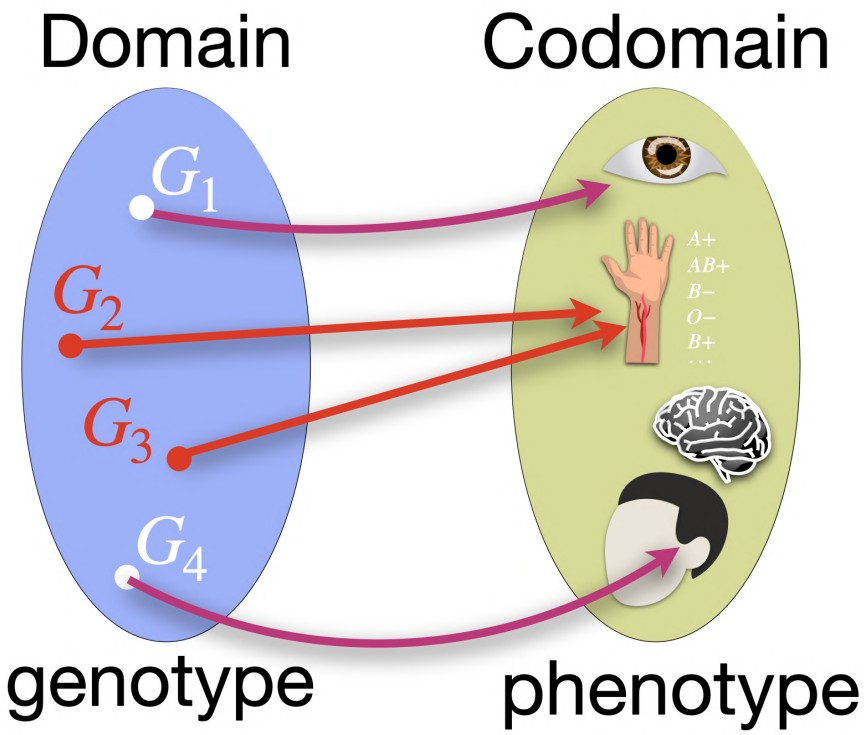
Your **phenotype**, though, is much more exciting. It is the manifestation of your genotype: the observable traits that result from your genes, like your eye color, your blood type, or the size of your brain!



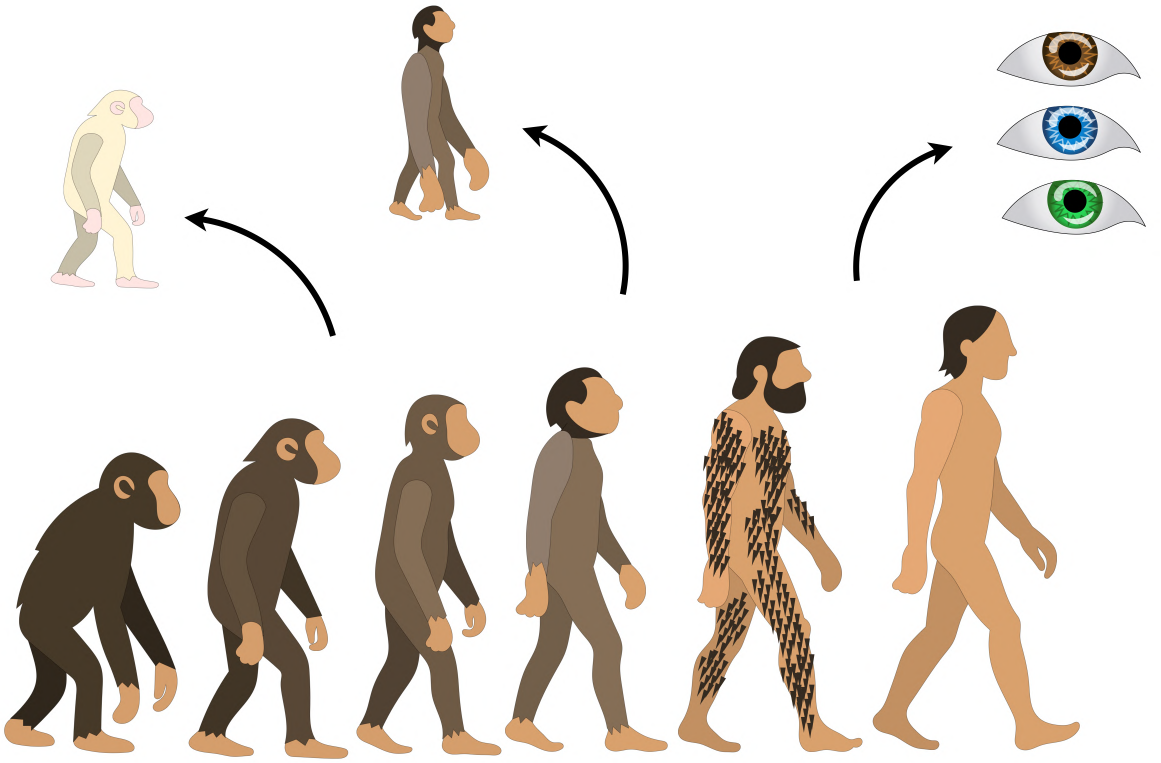
There is a mapping where the domain is your genotype and the codomain is your phenotype.



If the environment stays fixed around you (i.e. constant temperature, nutrition, stress, chemical exposure, and so on), this mapping behaves like a function, but it's not one-to-one, it is *not injective*.



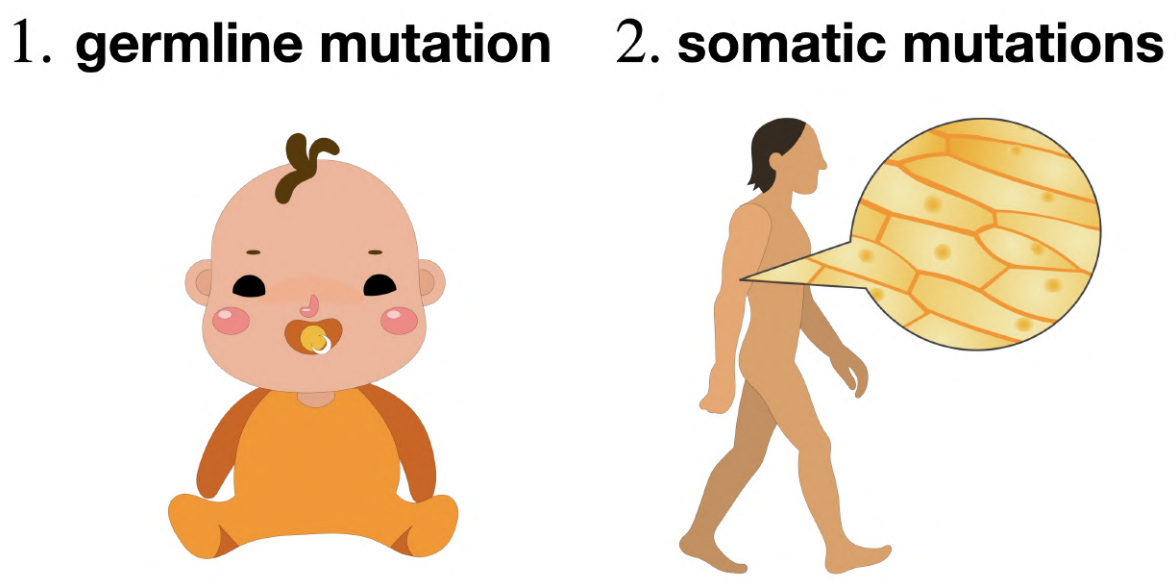
Many different genotypes can lead to the same phenotype. And that's actually a good thing, because it creates stability in biological evolution. It allows nature to explore variations without breaking what already works.



There are two main types of genetic mutations:

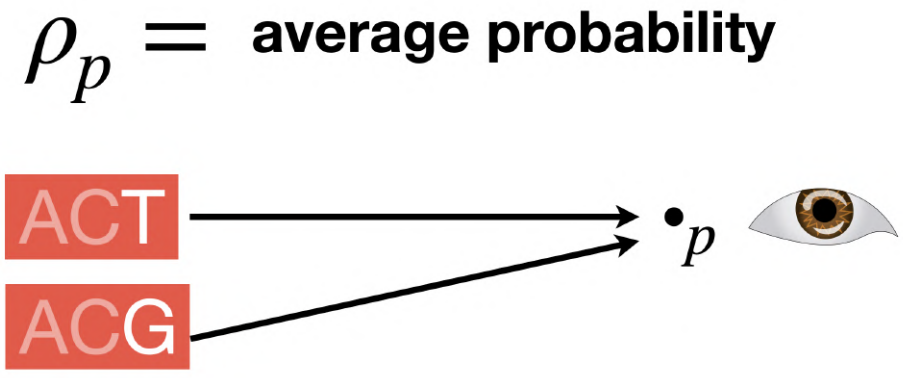
1. **Germline mutations**, which happen in sperm or egg cells, and can be passed on to children. So before you're born;

2. **Somatic mutations**, which happen in the body's cells usually during your lifetime.



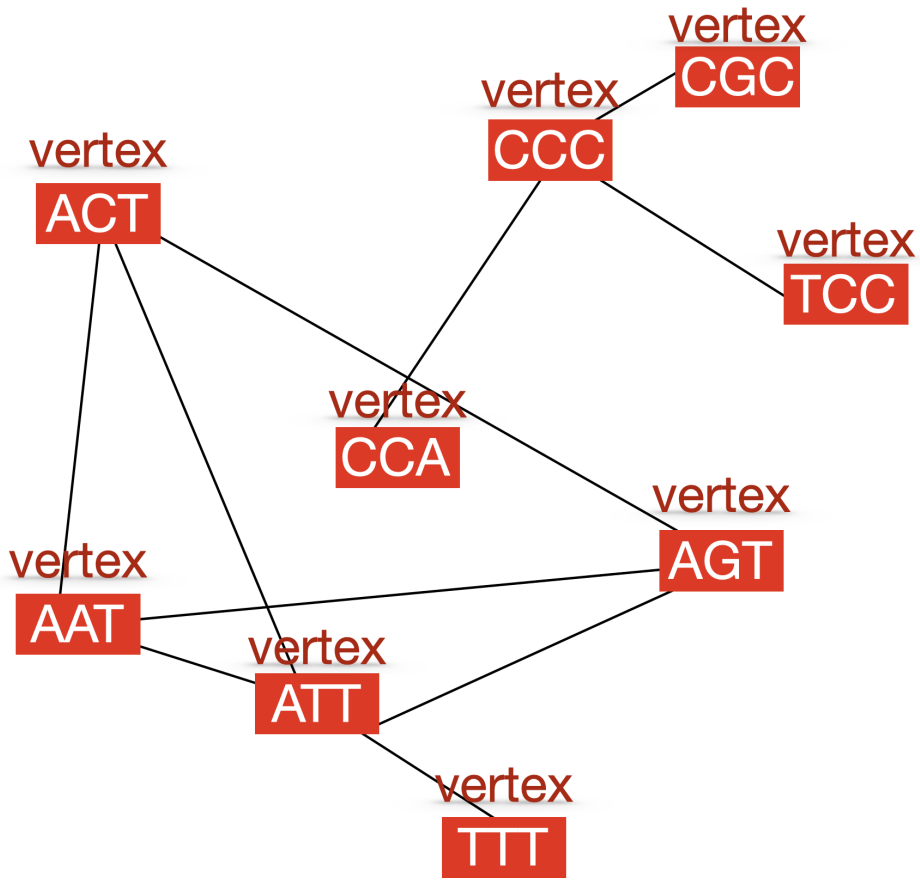
Somatic mutations are usually the dangerous ones. They can lead to diseases like cancer. But thankfully, our biology is built on a mathematical system that provides a kind of stability against mutations.

This tendency for a trait to stay stable despite mutations is called **phenotype robustness**. Mathematically, it can be expressed as the average probability ρ_p that a single-character mutation of a genotype mapping to a phenotype p does not change the phenotype p.



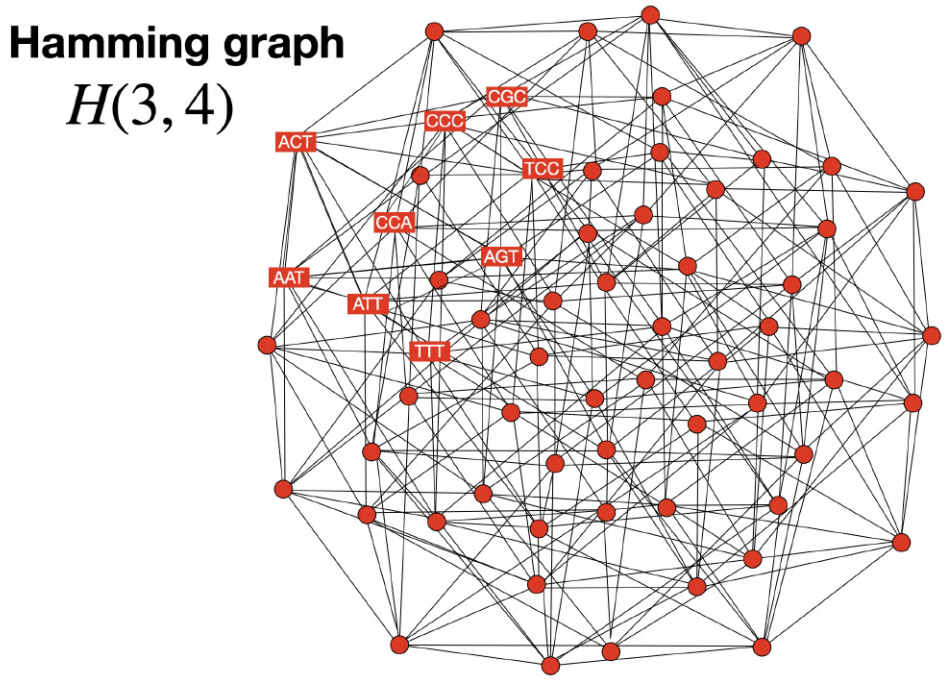
Phenotype robustness has a mathematical structure. Each genotype (i.e. a string of letters over the set $\{A, C, G, T\}$) is a *vertex*. Two genotypes are connected if they differ by a single mutation. In other

words, if they differ in exactly one character. This structure is called a *Hamming graph*: a high-dimensional graph where *edges* represent a single mutation.

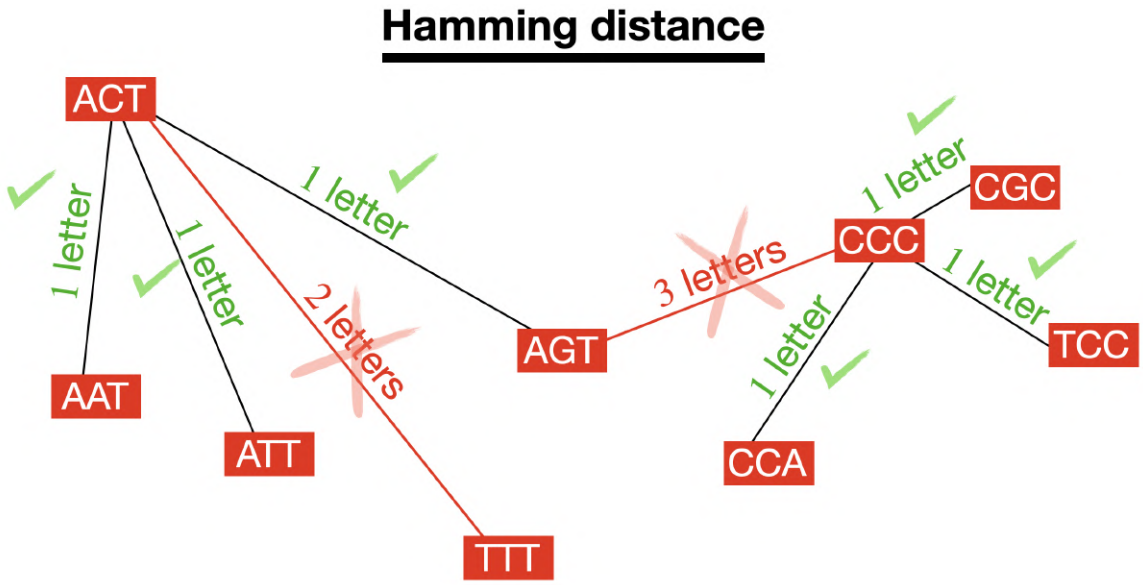


Let's see a concrete example to illustrate the point:

Take the Hamming graph $H(3, 4)$. Here, each vertex is a string of 3 letters over the alphabet $\{A, C, G, T\}$. The vertex set is written as $V = \{A, C, G, T\}^3$ (the Cartesian product of $\{A, C, G, T\}$ (the alphabet set) with itself 3 times), which means all possible combinations, with repetition, to form strings of length 3.



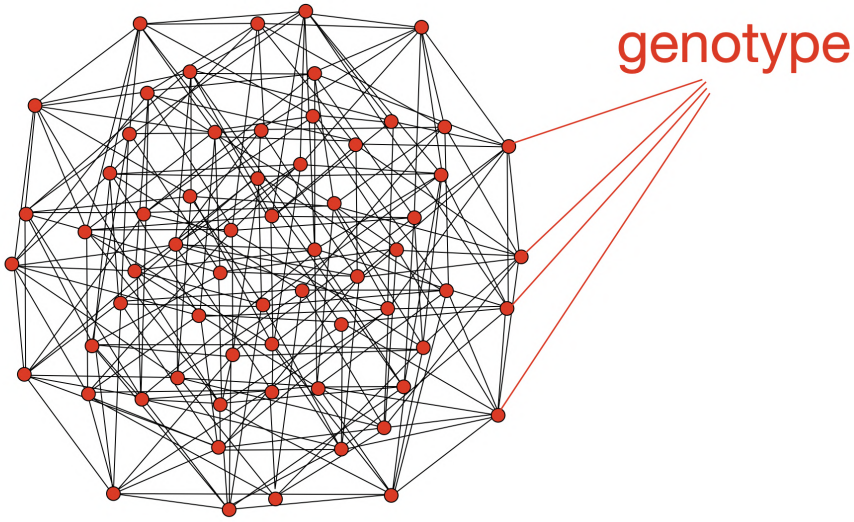
Two vertices are connected by an edge if they differ in exactly one position. That's what's called a Hamming distance of 1.



Using simple combinatorics, there are $\underline{4} \cdot \underline{4} \cdot \underline{4} = 4^3 = 64$ such strings. And thus this graph has 64 vertices, each one representing a possible genotype.

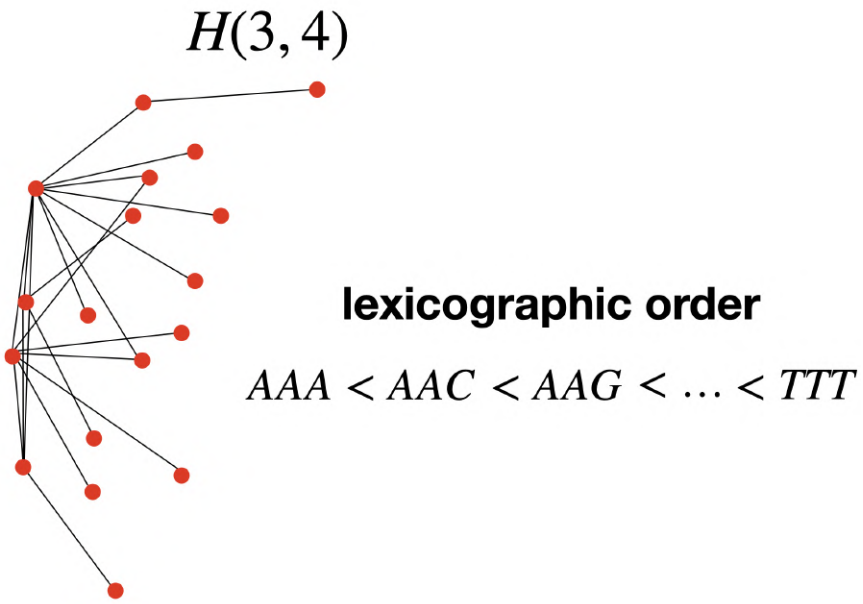
vertices = $\underline{4} \cdot \underline{4} \cdot \underline{4} = 4^3 = 64$

$H(3,4)$

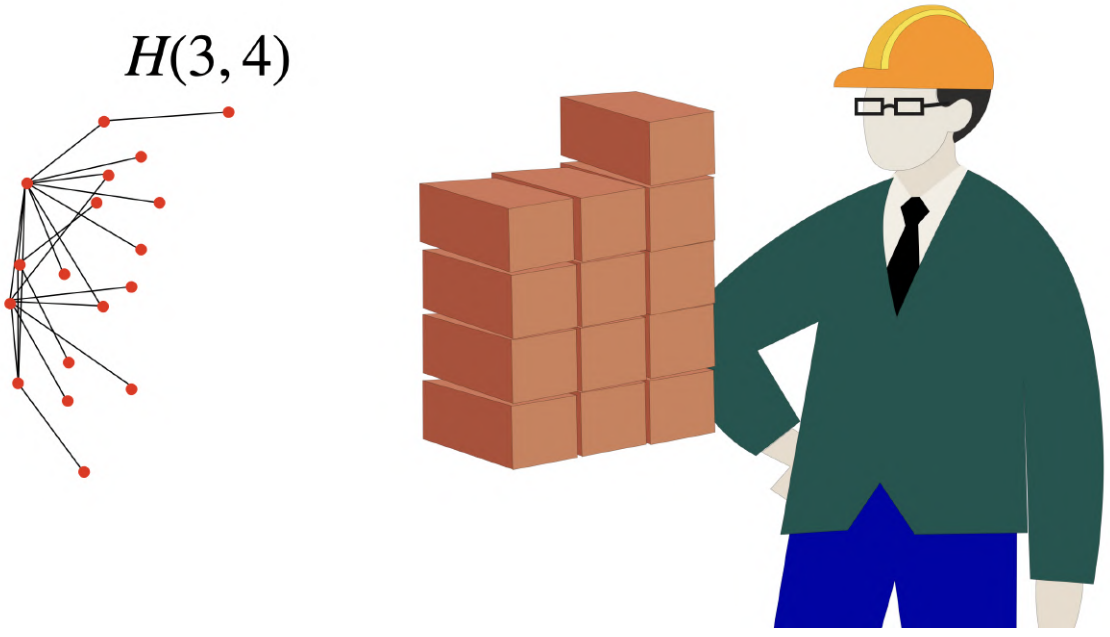


Note: *If you'd like to be the first to find out when we launch our very first books and courses, sign up with your email address on our homepage, dibeos.net.*

Now, imagine that we don't build the entire graph at once. Instead, we build it incrementally, vertex by vertex, following a rule called **lexicographic order**. In simple terms, that just means we order the genotype the same way words are ordered in a dictionary (comparing letters from left to right: $AAA < AAC < AAG < \ldots < TTT$).

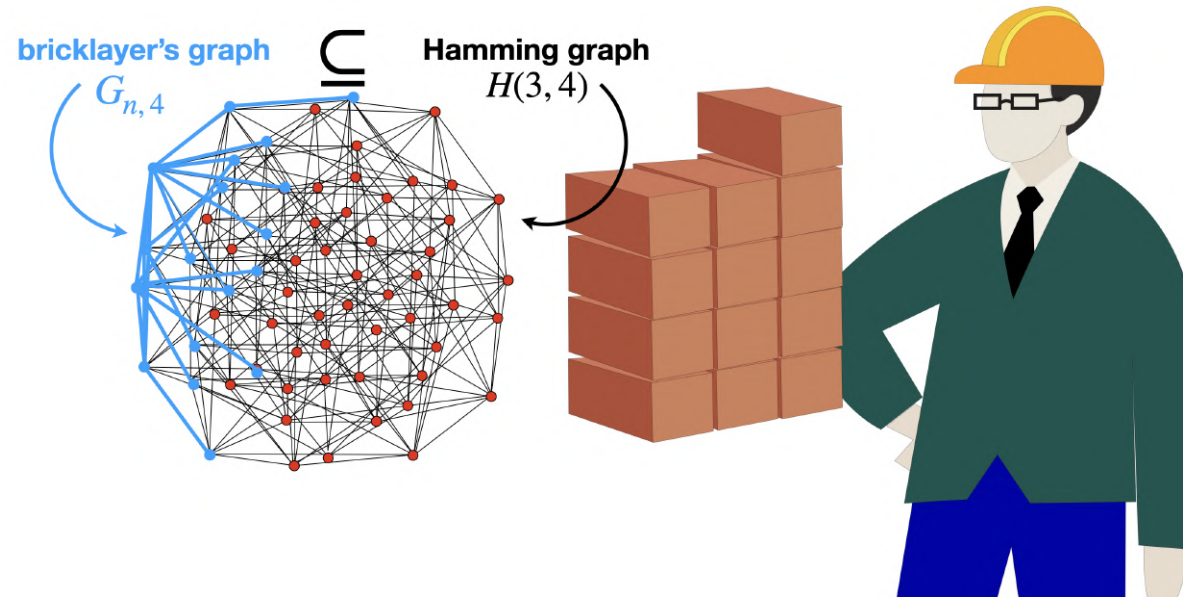


This way of adding vertices is like a bricklayer stacking bricks row by row, and that's why we call it a *bricklayer's subgraph* of the full Hamming graph.



At each step, we add a new genotype (a new vertex) and connect it to the previous ones if they differ by a single character. In this growing

structure, every added vertex reflects a new genetic possibility, and every edge is a mutational path.



Bricklayer's graphs are important in mathematics because they let us study how properties of the Hamming graph emerge step by step, and they surprisingly connect to number theory through the **sum of digits function**.

Fun Fact

As you can tell this is just starting to scratch the surface of what kinds of mathematics are involved in genes: graph theory, analysis, combinatorics, probability, number theory, and we can go on forever. I mean,

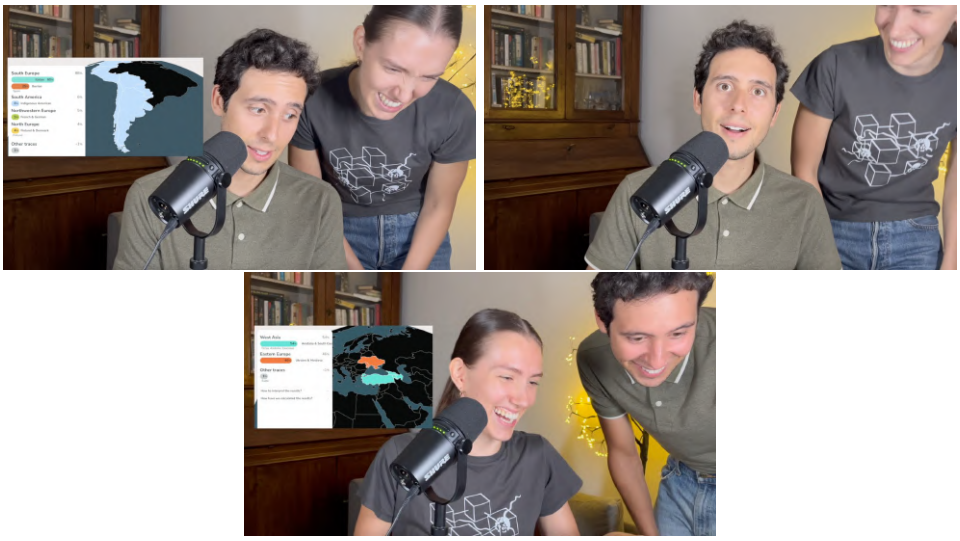
organisms are basically walking math.

But, wouldn't it be fun to actually know a bit more about your own genes? To see all of this math in action?

A couple months back we were approached by the company **ADNTRO**, who gave us the opportunity to know just that. Thankfully, there were no needles involved, just a container to spit into.

We really really loved this entire process and the concept behind it. A DNA test is something we've personally been wanting to do for a while, so we were super curious to see the results.

When we got our results back... well, we were shocked! (watch the video in our YouTube channel to see our reaction).



We saw that there's so much more you can learn than just ancestry. What's your perfect diet? Which vitamins should you be taking? How likely are you to get joint or muscle injuries? What diseases are you prone to or are more protected against? And that's just a small part of it. I mean there is just a treasure trove of knowledge inside of you. It even tells you if your genetics are predisposed to **math abilities**! Now how cool is that?

Once you've done your DNA test, **ADNTRO** always upgrades the features on their website so that you can stay up to date with new research. You can find out more and more about your genes as science moves forward, or as they upload new features.

If you've taken the test with another company, chances are they didn't give you some, if not most of these stats. BUT, don't worry, **ADNTRO** lets you upload your DNA results even if you've taken them with another company.

They also don't sell or share your data, and you can download it at any time and even delete it from their systems if you want to. So these results are completely yours.

Honestly, there's literally no better way to invest in yourself. Or, invest in someone else by getting them the test kit as a gift. If you'd like to find out more, click this link AND get 10% off by using the coupon code **DIBEOS** to make sure you get that discount.

Let's get back to the math now.

The Sum of Digits Function

The sum of digits function is actually very simple. For example, the digit sum of the decimal number 1995 would be $1+9+9+5=24$.

All you need in order to define this function is a number i (or sequence of digits) and a base b: $s_b(i)$

For example:

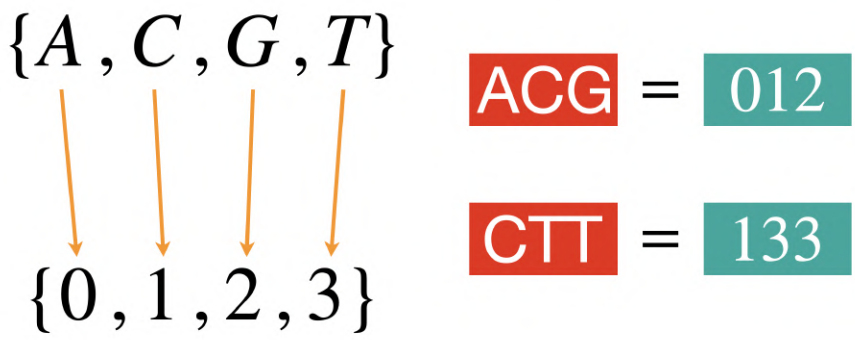
$$s_2(5) = 1+0+1 = 2$$

(because the base is $b = 2$*, so we need to write* 5 *in binary, which is* 101*)*

$$s_{10}(123) = 1+2+3 = 6$$

(because the base is $b = 10$*)*

The thing is, we don't actually have strings of digits in our DNA case study. Our genotypes are written with letters, like ACG or CTT. So before we can use the nice math of Hamming graphs and digit sums, we need a way to translate letters into numbers. The simplest trick is: just assign each letter of the genetic alphabet to a digit. For DNA, the alphabet has 4 letters, so we map them to the 4 digits $\{0, 1, 2, 3\}$.



Then the genotype ACG becomes the digit string 012, and CTT becomes 133. Now, we can treat every genotype as a base-4 number, and that makes it possible to use combinatorics and number theory on

them.

There is a theorem that says:

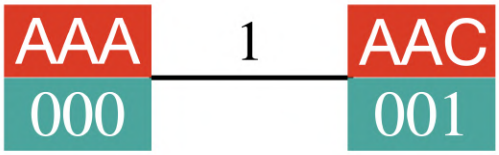
Theorem: Let $G_{n,k} \subset H(l,k)$ be a bricklayer's graph, which is a subgraph of the Hamming graph $H(l,k)$ of strings with length l over the alphabet of k letters, such that n is its the number of vertices. Then the number of edges is:

$$\boxed{|E| = S_k(n) = \sum_{i=0}^{n-1} s_k(i)}$$

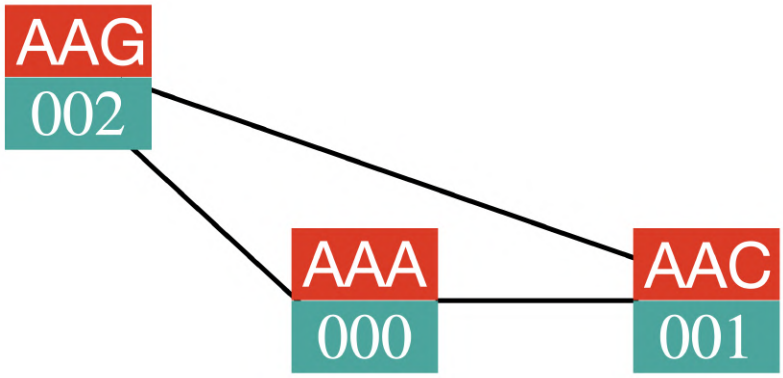
Where $s_k(i)$ (little s) here is the **sum of digits** of the base-k representation of i, and $S_k(n)$ (capital S) is the **cumulative sum of all those digit sums**.

Ok, there's a lot of information packed here. Let's see a concrete example to illustrate the beautiful mathematics behind it, and how this directly influences your DNA.

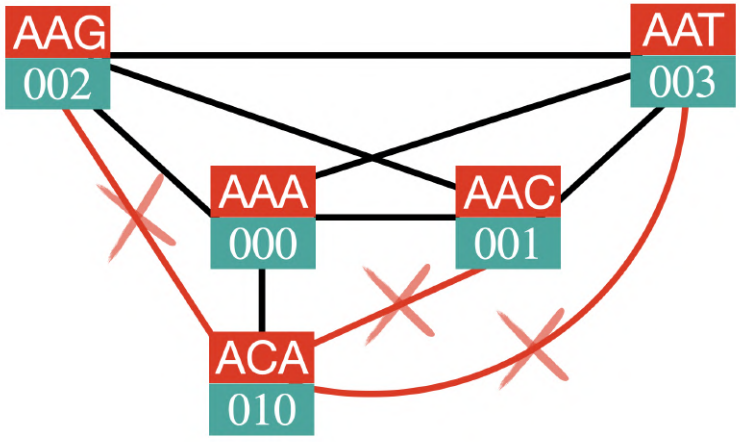
Since we need to follow the lexicographic order, we start with AAA, or 000. Next in line would be AAC, or 001. And we notice that we can already establish our first connection (or edge), since they differ by only one character. In fancy terminology, their Hamming distance is 1.



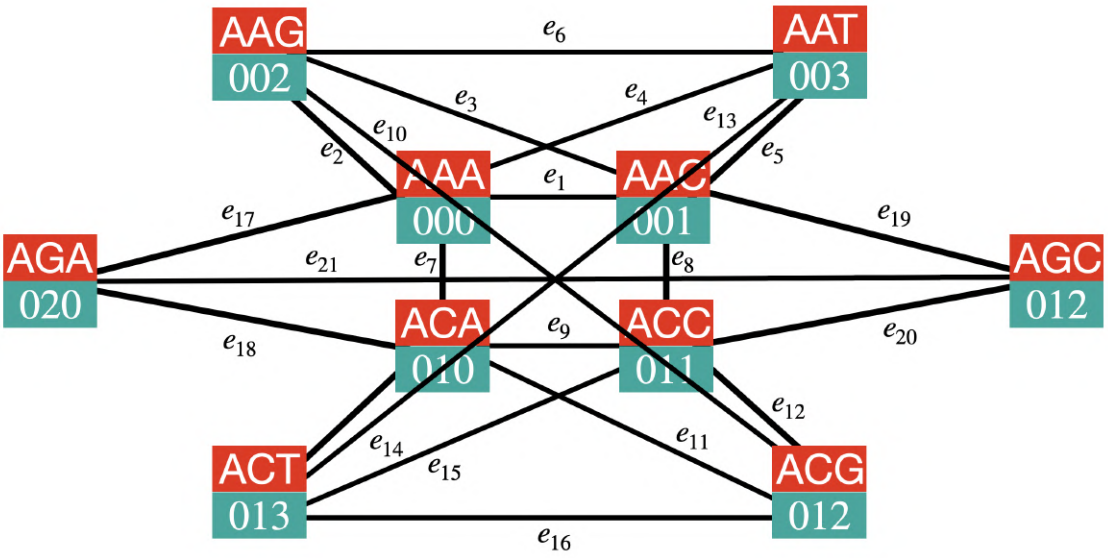
The next vertex is AAG, or 002, which has 2 edges (so far).



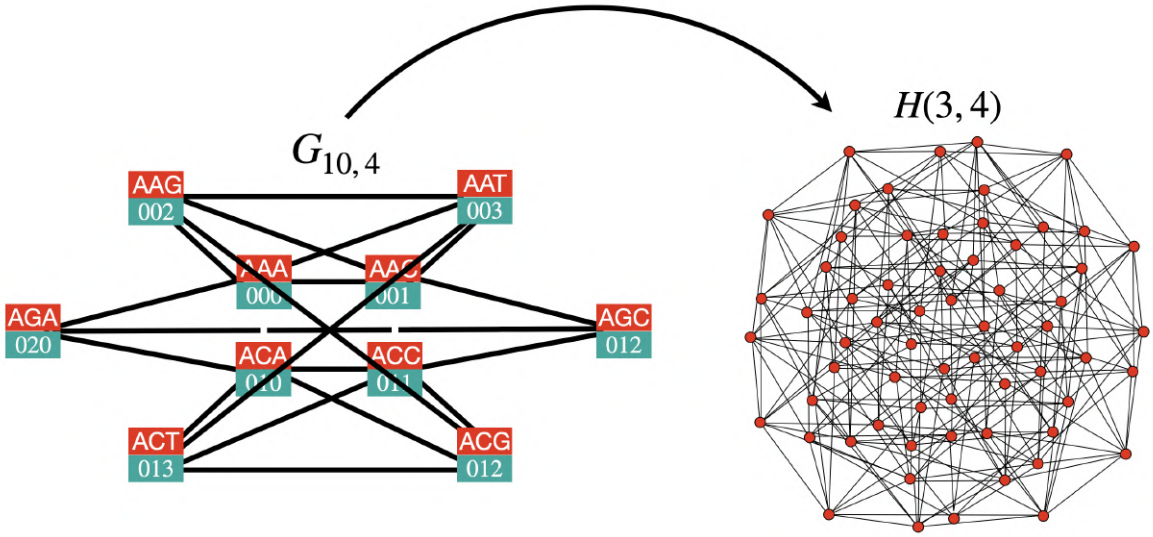
If we continue this process up to ACA, or 010, we notice that it has an edge only with AAA, or 000, but not with the others that come before it, since they differ by more than 1 character. However, it still can have more edges connecting it with the later genotypes in the sequence.



Let's go on and build this bricklayer's graph up to $n = 10$ vertices, and enumerate the edges.



Just reminding you that this is a truncated version of the full Hamming graph $H(3,4)$. We also notice that this bricklayer's subgraph $G_{10,4} \subset H(3,4)$ has a total of 21 edges.



Just for you guys to know, we counted all of them one at a time in order, which was not easy. But wait a second, why are we counting them this way if we already have a formula to calculate the number of edges

for any general bricklayer's graph $G_{n,k}$?

Let's just apply the formula instead:

($n = 10$, $k = 4$ and $i \in \{000, 001, 002, 003, 010, 011, 012, 013, 020, 021\}$)

$$\boxed{|E| = S_k(n) = \sum_{i=0}^{n-1} s_k(i)} \implies |E| = S_4(10) = \sum_{i=0}^{9} s_4(i) =$$

$$= s_4(0) + s_4(1) + ... + s_4(9) = (0+0+0) + (0+0+1) +$$

$$+(0+0+2) + (0+0+3) + (0+1+0) + (0+1+1) +$$

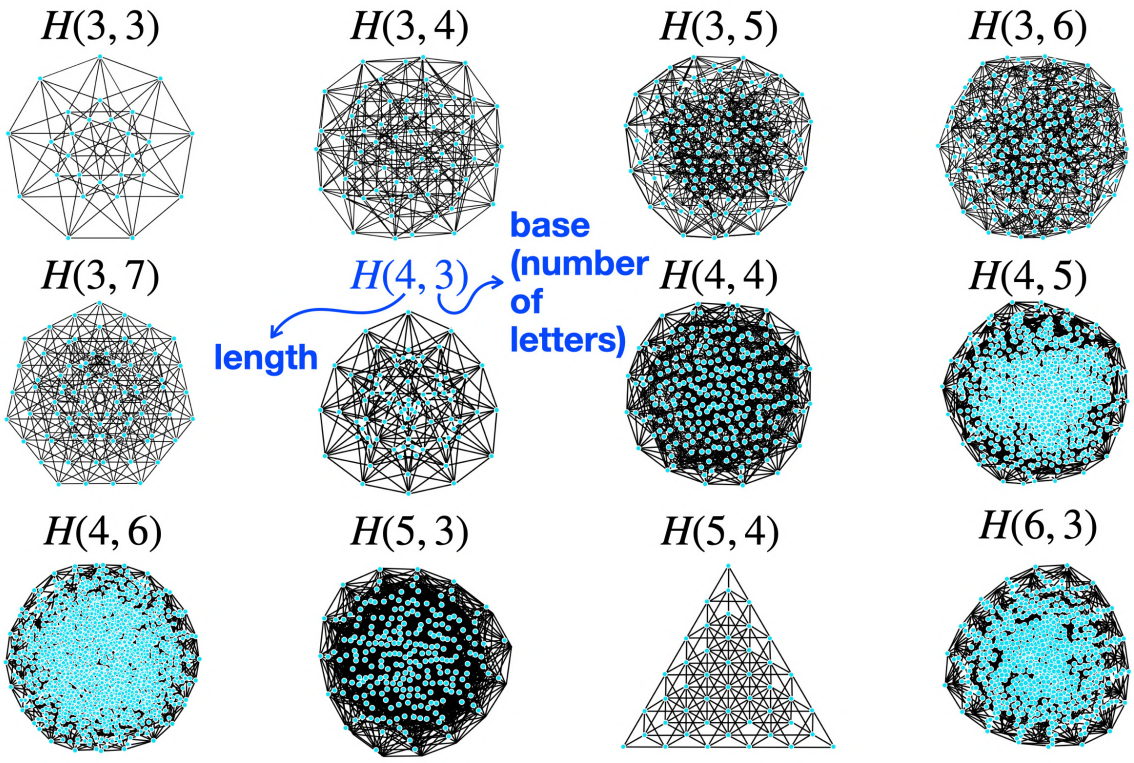
$$+(0+1+2) + (0+1+3) + (0+2+0) + (0+2+1) =$$

$$= 21$$

We got 21 edges! Just as before.

That's how this formula works.

Now, what happens when we also vary the number of letters k in our alphabet and the length of each string? We get many different possible graphs.



You might ask: *"ok, as a mathematical curiosity, altering the number of letters in the genetic alphabet to more than* 4 *is pretty interesting, but we know that we humans have only these* 4 *letters anyway. So, I guess this has no practical application to real life, right?"*

Wrong!

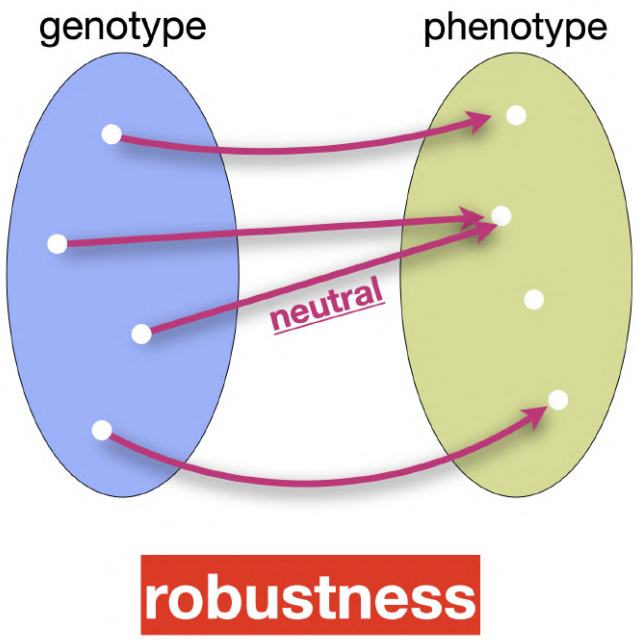
Scientists in synthetic biology have already engineered expanded alphabets. So, they've added new artificial DNA bases beyond A , C , G and T. These expanded alphabets can produce proteins with new amino acids, which allows us to explore entirely new forms of biochemistry. That means new drugs, new materials, and even new life forms designed in the lab. Not to mention that nothing prevents us from finding future extraterrestrial forms of life with more than 4 letters in their alphabet sets. We just don't know yet...

Nice, but the real question here is: how can we measure the phenotype robustness that we talked about earlier? In other words, how can we quantify this idea of stability despite mutations? So how can we measure how stable it is?

Well, first we need to define *robustness* mathematically.

Phenotype Robustness

Phenotype robustness (or simply *robustness*) is the probability that a mutation does not change the phenotype (i.e. is neutral).



We calculate it using this formula:

$$\rho\left(G_p\right) = \frac{2\left|E\left(G_p\right)\right|}{l(k-1)\left|V\left(G_p\right)\right|}$$

The numerator counts all neutral mutational possibilities (since each edge counts twice, once from each endpoint). The denominator counts all possible single mutations.

So this fraction is exactly the probability that a random single mutation has a neutral effect on your phenotype.

Let's see a quick illustration:

$(n = 10\ ,\ |E| = 21\ ,\ |V| = 10\ ,\ l = 3\ ,\ k = 4)$

$$\rho = \frac{2 \cdot 21}{3(4-1) \cdot 10} = \frac{42}{90} \approx 0.466$$

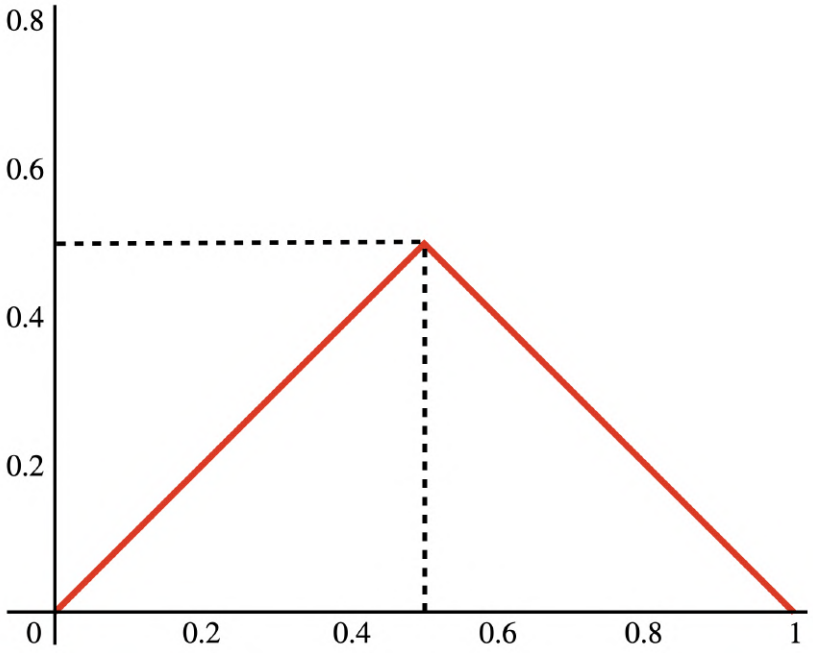
Therefore, for this small bricklayer's subgraph, on average, about 47% of single-letter mutations keep you inside the same phenotype.

Not super high, not super low. About half of mutations are completely neutral here.

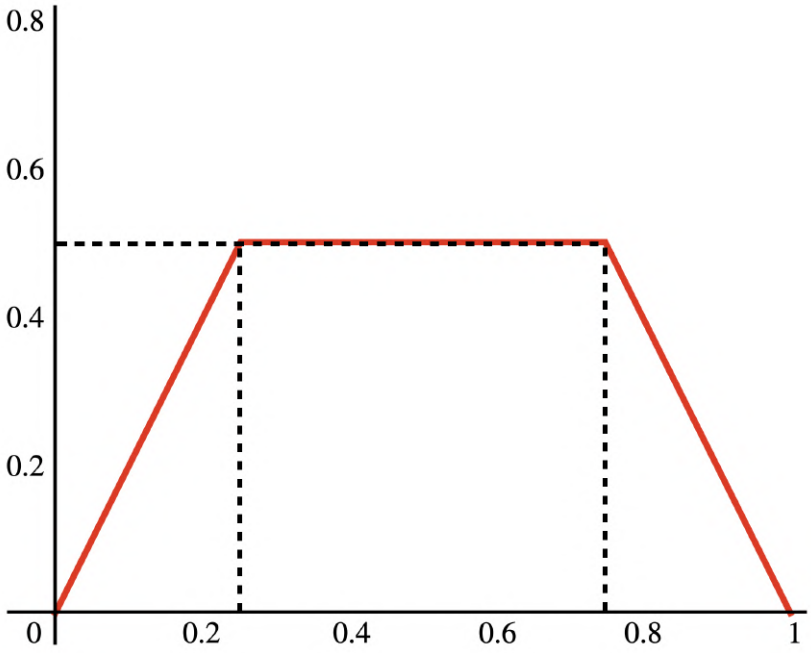
Now, let's talk about the *blancmange function* and how it relates to phenotype robustness.

The Blancmange Function

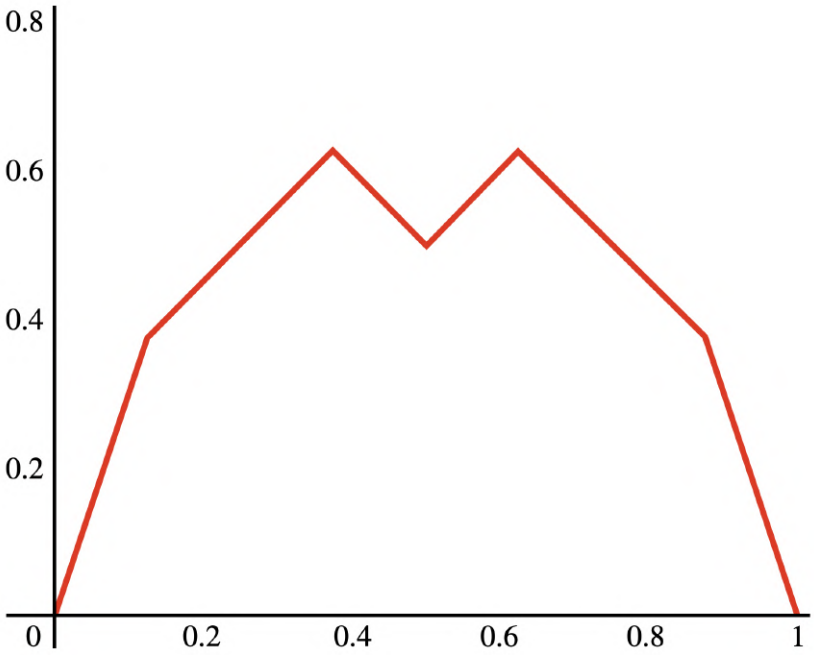
Plot a "triangle wave" function:



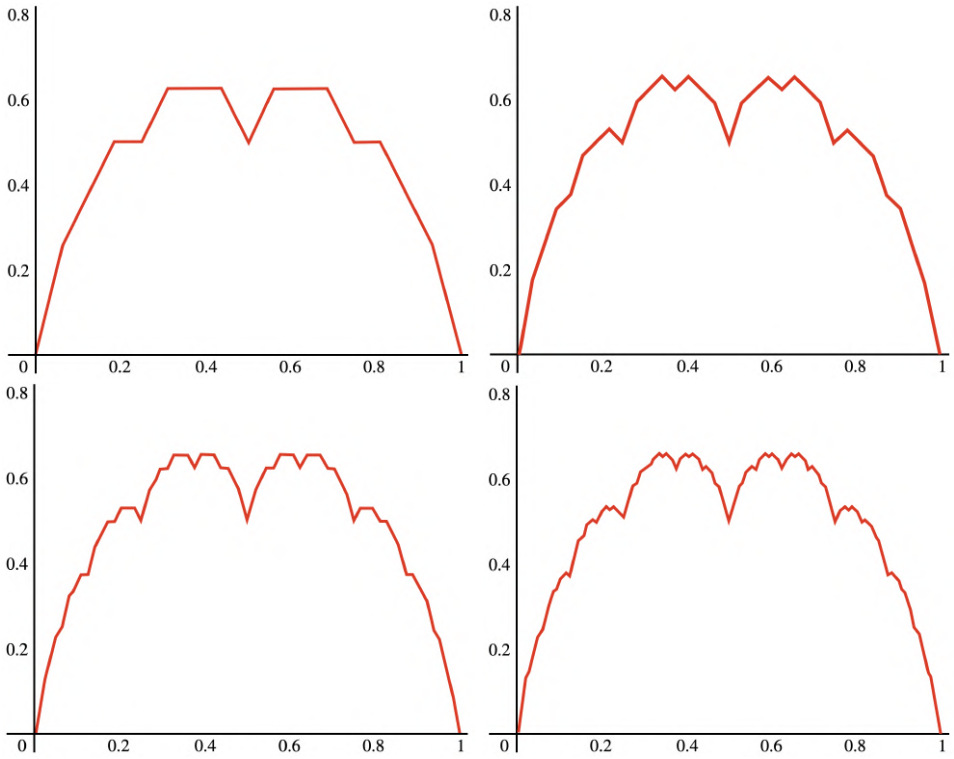
Then, for the interval $[0\,,\,0.5]$ create another sort of triangle (or "corner"), and do the same for the interval $[0.5\,,\,1]$ as well:



Then, we do it again for these 4 intervals ($[0\,,\,0.25]$, $[0.25\,,\,0.5]$, $[0.5\,,\,0.75]$ and $[0.75\,,\,1]$):



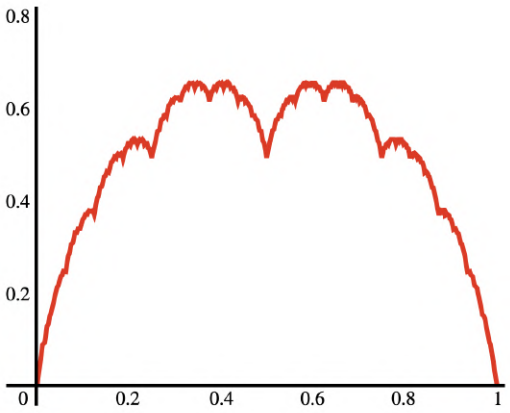
We can actually continue this process infinitely, and get a fractal that looks like a blancmange dessert.



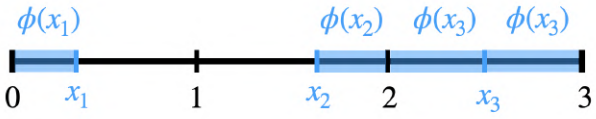
Interesting! The iterative formula for producing this fractal is

$$\boxed{f_N(x) = \sum_{j=0}^{N} \frac{1}{2^j} \phi(2^j x)}$$

Where $\phi(x)$ is the *distance of x to the nearest integer*.



$$f_N(x) = \sum_{j=0}^{N} \frac{1}{2^j} \underbrace{\phi(2^j x)}_{\phi(x)}$$



Let's see a few iterations using this formula:

$$f_3(0.5) = \sum_{j=0}^{3} \frac{1}{2^j} \phi\left(2^j \cdot 0.5\right) = \frac{1}{2^0} \phi\left(2^0 \cdot 0.5\right) + \frac{1}{2^1} \phi\left(2^1 \cdot 0.5\right) +$$

$$+ \frac{1}{2^2} \phi\left(2^2 \cdot 0.5\right) + \frac{1}{2^3} \phi\left(2^3 \cdot 0.5\right) =$$

$$= \phi\left(0.5\right) + \frac{1}{2} \phi\left(1\right) + \frac{1}{4} \phi\left(2\right) + \frac{1}{8} \phi\left(4\right)$$

Now, since $\phi($ **something** $)$ is the distance of **something** to the nearest integer, we have that:

- $\phi(0.5) = 0.5$ (distance between 0.5 and 0 = distance between 0.5 and 1)
- $\phi(1) = 0$ (distance between 1 and 1 is *zero*)
- $\phi(2) = 0$ (distance between 2 and 2 is *zero*)
- $\phi(4) = 0$ (distance between 4 and 4 is *zero*)

Going back to calculating $f_3(0.5)$:

$$f_3(0.5) = 0.5$$

Let's do the same for $f_4(0.25)$:

$$f_4(0.25) = \sum_{j=0}^{4} \frac{1}{2^j}\,\phi\left(2^j \cdot 0.25\right) = \frac{1}{2^0}\,\phi\left(2^0 \cdot 0.25\right) + \frac{1}{2^1}\,\phi\left(2^1 \cdot 0.25\right) +$$

$$+\frac{1}{2^2}\,\phi\left(2^2 \cdot 0.25\right) + \frac{1}{2^3}\,\phi\left(2^3 \cdot 0.25\right) + \frac{1}{2^4}\,\phi\left(2^4 \cdot 0.25\right) =$$

$$= \phi\left(0.25\right) + \frac{1}{2}\,\phi\left(0.5\right) + \frac{1}{4}\,\phi\left(1\right) + \frac{1}{8}\,\phi\left(2\right) + \frac{1}{16}\,\phi\left(4\right)$$

Now, since $\phi($ **something** $)$ is the distance of **something** to the nearest integer, we have that:

- $\phi(0.25) = 0.25$ (distance between 0.25 and 0 is 0.25)
- $\phi(0.5) = 0.5$ (distance between 0.5 and 0 = distance between 0.5 and 1)
- $\phi(1) = 0$ (distance between 1 and 1 is *zero*)
- $\phi(2) = 0$ (distance between 2 and 2 is *zero*)
- $\phi(4) = 0$ (distance between 4 and 4 is *zero*)

Going back to calculating $f_4(0.25)$:

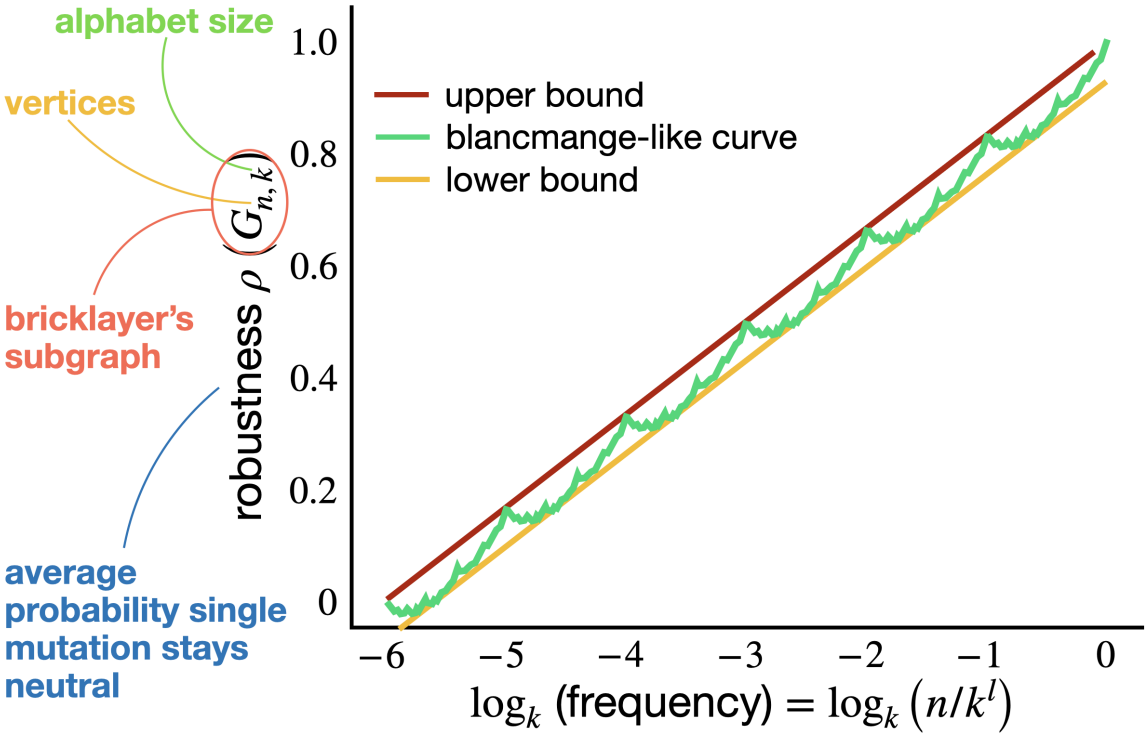
$$f_4(0.25) = 0.25 + 0.5 = 0.75$$

And so on...

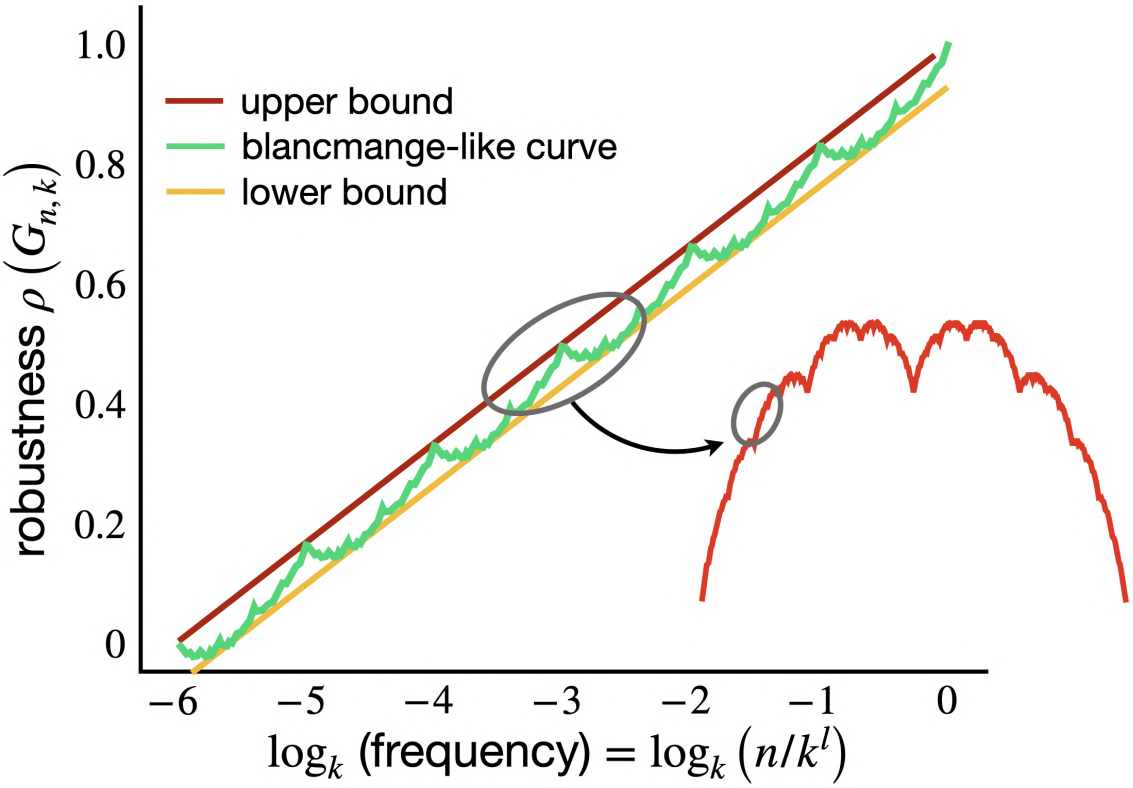
This is very cool, but you might be asking yourself right now: *"What does it have to do with anything we've seen so far?!"*

And you'll be surprised, because this is the **highlight** of this research!!!

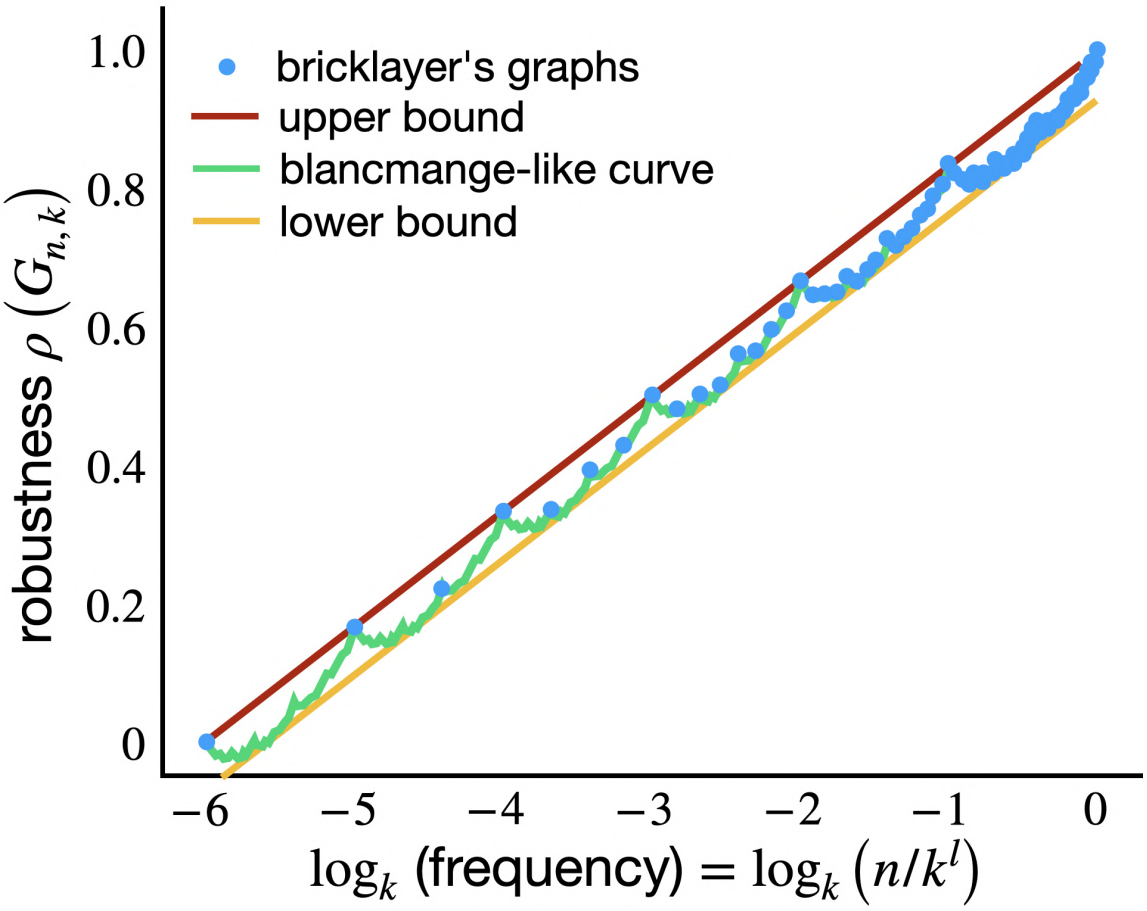
What happens if we calculate the phenotype robustness $\rho(G_{n,k})$, i.e. the average probability that a single-letter mutation stays neutral for a phenotype, across the bricklayer's subgraph $G_{n,k}$ (with alphabet size k and n vertices), and then plot it against the normalized frequency of genotype, given by $\log_k(\frac{n}{k^l})$?



This plot shows how robustness evolves as we add vertices one by one to the bricklayer's graph, and the surprising result is that the curve follows a fractal, the blancmange-like function, squeezed between upper and lower bounds.



There are of course many more technical details behind this study, and if you want to dive deeper you should check out the full paper linked in the description. The authors carefully worked out upper and lower bounds, the exact formulas, and biological interpretations of these graphs.



But here, our focus is on the mathematical result: contrary to all intuition, the plot of robustness doesn't form a smooth curve. It instead gives us a fractal structure (so a curve that is continuous everywhere, but differentiable nowhere). More specifically, a blancmange-like curve emerges "out of nowhere", which shows that mutational stability in genotype-phenotype maps follows the geometry of a fractal squeezed between very precise upper and lower bounds.

Crazy how the most abstract mathematics keeps on appearing in the weirdest places, huh?!

If you found this document useful let us know. If you found typos or things to improve, let us know as well. Your feedback

is very important to us. We're working hard to deliver the best material possible. Contact us at: dibeos.contact@gmail.com