Solution: ‘A Ticking Evolutionary Clock’

How to calculate when a gene’s rate of evolution will slow to a crawl.

By Pradeep Mutalik

Carrie Arnold’s article, “Evolution Runs Faster on Short Timescales,” explored new research showing that genetic changes that are quite brisk when measured over a few generations seem to slow down considerably when measured over millions of years. Our March puzzle attempted to see if such a result could be reproduced in a simple hypothetical DNA sequence.

Question 1:

> Imagine a gene that is 108 letters with A, T, G, C in random sequence. Assume that every year, there is a random change — one of the letters somewhere on this gene mutates and is replaced by one of the other three. After each year, you compare the current copy of the gene with the original and tally how many letters have changed. After a certain time “the evolutionary clock will have slowed to a crawl” — that is, the number of changed letters will have stopped rising. The evolutionary rate from here on is zero. How many letters of the original gene will have changed at that point? How many years will it take to get to this point? Is the curve exponential?

As the letters of the gene change, a point is reached when the probability of a preserved original letter (u) mutating to a different one is the same as an already mutated letter (m) changing back to the original one. Any change in a u converts it to an m, whereas an m has only a one-third chance of changing back to a u. So at a point when there are three times as many m’s as there are u’s, the evolutionary rate will become 0. This happens when there are 81 mutated letters (three-fourths of 108).

The second part of this question can be answered in several ways, depending on how you define the endpoint. In their comments, readers obtained different answers for just this reason. Let’s consider a few of these endpoint choices.

One point of view looks at the expected, or average, number of mutated letters. Since there is always a finite, albeit diminishing, probability that some particular letter, just by chance, remains unchanged for thousands or even millions of years, the expected number of mutations approaches ever closer to 81 over time in an asymptotic manner but never quite reaches this value. So one way to proceed is to define an endpoint of 80.5 as Michael did. This approximation is based on the fact that, at a point this close to equilibrium, there is a 50 percent chance that the number of mutations will increase the next year and push the total number over 81. The endpoint in this case takes 410

years.

Incidentally, this quick method of using the average expectation up to a point close to the desired value, and then using the fluctuations to "go over the top," brings to mind the following elementary school puzzle: A snail is at the bottom of a 20-foot well. Every day, it climbs upward 2 feet during the day, but slips back 1 foot during the night. After how many days of doing this will it be out of the well?

If you came up with 20, think again! The actual answer (click here to reveal) Here too, we use the average up to a point, and then rely on the discrete fluctuations to get over the top.

For more accurate methods, several readers turned to the mathematics of Markov chains or Markov processes: In such processes, as in gene evolution, there is a chain of nodes whose state at a given time completely determines the possibilities available to it in the future , without the need for a prior history. There are sophisticated mathematical techniques involving things like "finding eigenvalues of transition matrices" that mathematicians use to deal with Markov processes. But that does not mean that the ideas behind these processes are difficult to grasp — they just involve the repeated applications of some simple formulas. In the present problem, you can easily model the expected value of the gene's mutations in a spreadsheet using two columns with two simple repeated formulas as described below (you can do it with a single column, but the two-column version is a little easier to understand). In column A, put the numbers of consecutive years starting from 0 in cell A2, going down the rows all the way to 1000 at A1002. Label cell B1 in Column B "Original" and cell C1 in Column C "Mutated." In year 0, we start with 108 original letters and no mutated ones, so put 108 in cell B2 and 0 in cell C2. In the next row we will place the two formulas that describe how the numbers change from year to year. The number of original letters can decrease by a mutation that strikes an original letter from the previous year (probability = u/108) or can increase by a mutated letter reverting back to the original, which will happen a third of the time (probability = (m/108)/3). This is captured by the following formula in cell B3: = *B2-(B2/108)+(C2/108)/3*. Similarly, the change in the number of mutated letters based on the previous year is given by the following formula in cell C3: = *C2-(C2/108)/3) + B2/108*. These formulas can be copied over all the 1,000 rows and voilà, you can see how the number of expected mutations changes, inching ever closer to 81 as the years go by, passing 80.5 at 410 years.

The two equations shown above can be combined with some algebra into a single iterated function: $m_n = (80/81)m_{n-1} + 1$ where m_n is the expected number of mutated genes in year n. This yields the following exponential formula, described by Mark P: $m_n = 81 - 81*(80/81)^n$. This is an exact exponential, which covers half the remaining distance to 81 mutations about every 56 years. The number of mutations starts off by increasing rapidly, then tails off and slows to a crawl.

A second approach to solving this problem, rather than looking at the expected values, is called the "hitting time," and is standard in Markov processes. Here, the number 81, the dynamic equilibrium point, is treated as the "absorbing node." This means that whenever any gene reaches 81 mutations, it is considered to have hit its target and is "absorbed," and the future evolution of this gene is not considered. Using this approach, Mark P. got an answer of 262 years, while Ashish and Dennis obtained 281.8 years.

The reason for the difference can be explained as follows. Imagine that there are, say, 1000 separate isolated populations that all have the same evolving gene, with the same mutation rate. The question that Mark P. answered is: How many years will it take for 50 percent of these gene populations to reach 81 mutations? This is the median value. Ashish and Dennis, on the other hand, found the average time for all the 1000 genes to reach 81 mutations. This time is slightly longer because there will be some genes that will take far longer times just by chance. The spreadsheet method that I

described above can be used to obtain both these answers. It just takes a larger spreadsheet — one that has 82 columns, one for each number of possible mutated letters from 0 to 81, again having simple formulas relating the number of mutations from one year to the next.

Note that the “hitting time” approach gives a smaller number than the expected value of about 400 years that we calculated earlier. This is because the hitting method does not take into account the fact that the genes keep changing even after reaching 81 mutations, and some may dip below this number by random reversals, even after initially hitting the number earlier. So the hitting approach overestimates the actual proportion of genes with a given number of mutations in any year. If continued evolution is included in the model, as can be done by extending the humble spreadsheet approach to 109 columns, you can see that the first time that 50 percent of the genes actually have 81 or more mutations in a given year is closer to the expected value we calculated earlier — about 400 years.

Question 2:

The above scenario is not very realistic. Every letter in a real-life gene sequence has a different chance of having a mutation that “sticks.” The letters at some locations in the DNA sequence are preserved, because changes in them are catastrophic; others, at inconsequential locations, can change readily. One general rule is that the third letter of every triplet can change easily. This is because the third position in a triplet is often redundant: The first two positions fix the amino acid the triplet codes for.

Assume that the third letter of each triplet is three times as likely to get mutated as are the first and second. Now try to answer the same questions as in Question 1.

The rate of evolution is slower here, as is to be expected. There are 36 letters that change three times as rapidly as the other 72. So a mutation in a given year has a 3-in-5 chance of striking the third letter of a triplet, and only a 1-in-5 chance of striking the first or second letters. The dynamic equilibrium point is still the same as before: 81 mutations. The number of years to 80.5 mutations using the expected value approach is 630 years. Mark P.’s endpoint is reached in 432 years, and Ashish and Dennis’s in 579 years.

The solution methods are similar to, though somewhat more complex than, those for the first problem. The evolutionary change curve is now the sum of two different exponential curves, a fast-changing one that covers half the remaining distance to the equilibrium point every 31 years, and a slower one that does so every 93 years. The resulting graph still looks like an exponential curve, though it is no longer as good a fit to the exponential function as in the first case.

Bonus Question:

I gave this hypothetical piece of DNA 108 letters. What was my reason for choosing that number? Is it because 108 has mystical significance, as a Google search will indicate?

Mark P. came closest with this answer:

> The author could have just as easily chosen 36 or 12 or 8 to be the gene length for this question. The key here is that with a setup like this (four possible letters, and thus three possible letters to change to), you need a number divisible by 4 so that 3/4ths of gene length is still an integer.

> But 8 is unsatisfactory because it doesn't lend itself to Question 2, where we divide everything up into triplets. (Yay for biological inspiration!)
>
> 12 would've satisfied all these constraints, as would've 36. But both those numbers sound unrealistically small for the length of gene. Maybe that's why the author chose 108.

Well done, Mark. That's almost exactly correct! Actually, even 108 is much too small a number for the length of a gene: Even the smallest genes have several hundred bases. I wanted a number around 100 so that the calculations wouldn't become as complex as they would for a real gene.

In fact, as I mentioned and as Barbara West stressed, this simple model is far from biologically realistic. Most proteins need to be 70 to 80 percent preserved in order to execute their functions. This is the reverse of our mathematical example in which the change is 70 to 80 percent, and only 20 to 30 percent of the original gene is preserved! Another point is that the actual mutation rate needs to be slower by a factor of hundreds or thousands to be realistic.

This drives home the point that while mathematical models can be useful in biology, realistic ones are unbelievably messy and have to keep track of many parameters and exceptions. We saw that even in our simple example, the fit to simple mathematical functions loosened after we added some biological realism. Making accurate models requires the math to follow the biology and not vice versa. And when conclusions are drawn from mathematical models, it is best to use them as general guides without treating them as sacrosanct and following their precision over and above what the assumptions, simplifications and the data warrant. For our puzzle, we just sought an exponential curve for the evolutionary rate that explained the finding of fast genetic change at smaller timescales compared to longer ones. We found several possible curves, depending on the endpoint selected: The actual numbers are not very significant. To make them truly usable, we will need really accurate mutation-rate and gene-preservation data from the field, among other things.

The *Quanta* T-shirt for this column goes to Mark P. Congratulations! Thanks to all who contributed, and see you soon with new insights.