Bioinformatics
MAGAZINE

November 5, 2023

Author: Alexandre de Carvalho Assunção

How Bayes's Theorem Can Help Us Make Vaccines?

Let's say someone somewhere ate something weird and got sick—sneezes, coughs, fever, body aches—the works. And then his family gets sick, and his coworkers, neighbors…you know where this is going, right? We got an outbreak. And let's assume it's a viral outbreak since everyone not wearing diapers has experience with those.

Let's also assume this mystery virus has a protein in its surface that binds to specific receptors in human cell membranes and this binding triggers the cell to engulf the virus, allowing it to go about its nefarious business replicating and murdering its host.

This protein would make a juicy target for a vaccine. If we could make a strand of mRNA that encodes it and inject it into our cells, it would be translated into the viral protein, recognized as a threat by the immune system, who would produce antibodies and *remember* this protein. If we're ever invaded by the real virus, its response would be quick and the virus would have far more trouble replicating inside of us.

But how do we go about making this strand of mRNA?

Well, the first step is to get a sample of the virus and sequence its genome, which could be made of DNA or RNA. Continuing with our theme of nostalgia, let's go with RNA. So someone in a white coat does all that fancy lab work and uses a high-tech machine that takes in a vial of virus-containing solution and spits out a file containing a string of roughly thirty thousand nucleotides.

The second thing we do is search every genome database to see if we are dealing with a known virus. We're not. No one has ever been crazy enough to eat whatever the dude in the first paragraph ate. This is a new virus. And our best bet to develop a vaccine is finding the gene encoding the protein the virus uses to invade our cells in its genome.

However, just because this is a new virus, it doesn't mean that it doesn't have anything in common with the thousands and thousands of other viruses in our databases, and there's plenty we can learn from them: we can, for instance, see what an average viral gene looks like. What is the average length? Are there any motifs that appear more often in genes instead of non-coding areas? In sum: if we observe a certain RNA sequence AUAC…CUA, what is the likelihood that it is part of a gene? We could also use the similar viruses in our databases to estimate the overall probability of finding a gene in a viral genome.

With these pieces in place, let's introduce some notation. Let's denote by H our hypothesis that the stretch of RNA we're analyzing is part of a gene, and by D the set of nucleotides we're observing.

Bayes's Theorem states that:

$$P(H|D) \ = \ \frac{P(D|H)P(H)}{P(D)}.$$

This theorem is trivial to prove:

$$By\ definition,\ P(A \cap B) \ = \frac{P(A|B)P(A)}{P(B)}.$$

$$But\ P(A \cap B) = P(B \cap A) = \frac{P(B|A)P(B)}{P(A)} \Rightarrow P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

But don't be fooled by its simplicity. Bayes's Theorem may just be one of the most important and powerful theorems in all of mathematics. Notice: we begin with a preconception of what the probability of finding a gene is, which we call our *prior*, $P(H)$. But then we observed some data—the nucleotides—and we updated our belief of whether we are analyzing a gene or not. This is a rational and systematic way to update beliefs based on observations.

Bayesian reasoning can be applied to all areas of science and life, but discussing that goes beyond the scope of this text. Let's go back to making our vaccine before the outbreak turns into a full-blown pandemic.

Let's use Bayes's Theorem to construct a model that can tell us whether we are reading a gene or not. When we scoured genetic databases looking for similar viruses, we calculated the frequencies each nucleotide appears in genes and in non-coding areas. We also estimated how much of a virus' genome is made of genes and the average length of a viral gene.

This is what we found:

If we're in a gene (G): $\{A = 30\%, C = 25\%, G = 30\%, U = 15\%\}$.
In non-coding areas (NCA): $\{A = 25\%, C = 10\%, G = 35\%, U = 30\%\}$.

We also find that around 80% of similar virus' RNA is made of genes, leaving 20% of it in non-coding regions. Lastly, the average size of a virus' gene is 1500 base pairs (bp's).

Given our mystery virus's genome $G = (ACACAGCCAUA...AUAC)$, we we have everything we need to determine the probability of position i being part of a gene.

Finding that 30% of viral genes is made of A's can be written as $P(A \mid GENE) = 0.30$. We call that our *likelihood*. Knowing 80% of the genome is made of genes gives us $P(GENE) = 0.8$. We call that the *prior*. The total probability of observing A is the probability of observing A given that we're in a gene times the probability we're in a gene plus the probability of observing A given we're **not** in a gene times the probability we're **not** in a gene. We call that the *marginal* and write it as:

$$P(A) = P(A \mid GENE)P(GENE) + P(A \mid NOT\ GENE)P(NOT\ GENE).$$

By Bayes's Theorem:

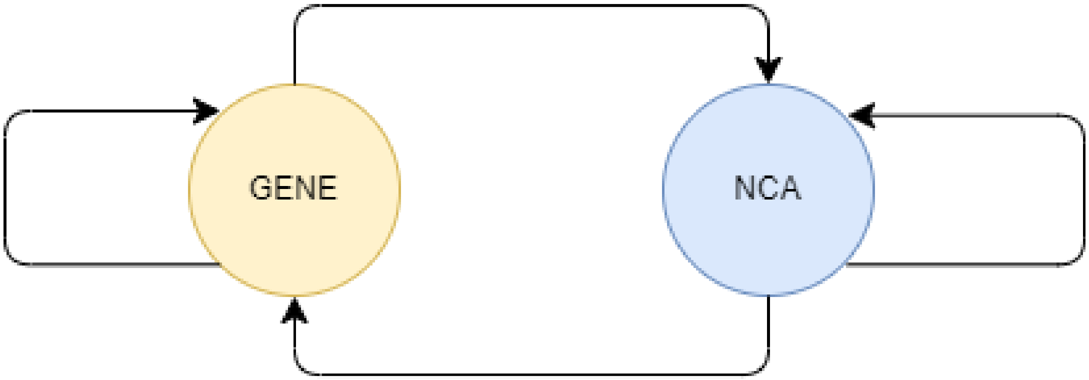
$$P(GENE \mid A) = \frac{P(A \mid GENE)P(GENE)}{P(A)}.$$

If I’m not botching this up too badly, you probably have some idea of how we use Bayesian Inference to find a gene in a full genome. But you’re also probably noticing that this model of inference is too simple. It assumes the state of every position of the genome is independent of every other, which we know is not true—coding a single amino acid requires three nucleotides.

We need a way to incorporate the genome’s sequential nature into our model—knowing the genome i position influences what we believe we’ll find in the $(i + 1)$ position.

So how do we incorporate this information to improve our model?

Think of it like this: at every position i, the genome can be in two states: a gene (G) or a non-coding area (NCA). And at the $i + 1$ position, we could stay in the state we were in the i position or change.



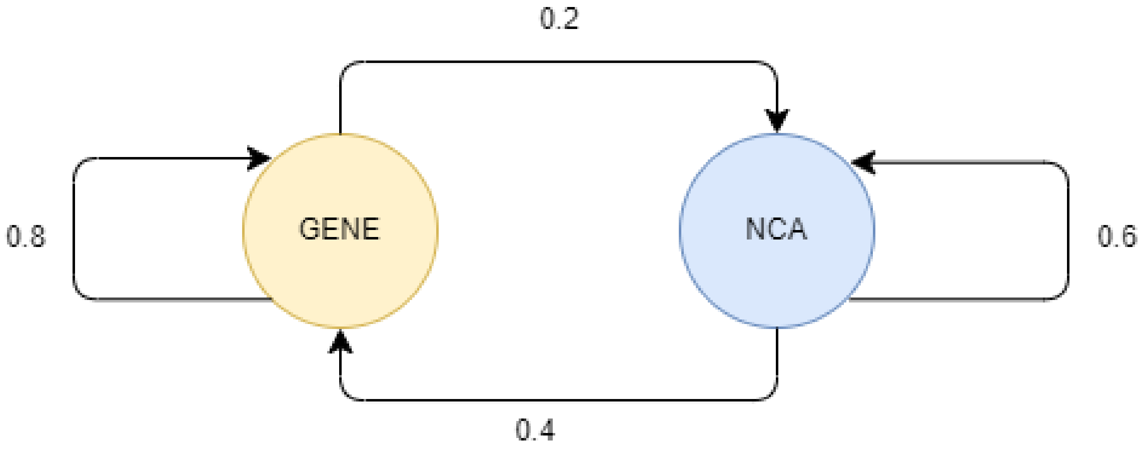
Now let’s think about how we “move” in this model. Suppose that at position i we’re at gene. We calculated the average size of a viral genome when we searched our databases and we can use that information to determine the probability of either staying at GENE in the next position or moving to NCA. We do that by imagining we're flipping a biased coin with sides GENE and NCA. The probability of staying at gene is:

$$\frac{1}{1 - AVERAGE\ LENGTH\ OF\ AN\ NCA}.$$

And the probability of switching to an NCA is:

$$1 - \frac{1}{1 - AVERAGE\ LENGTH\ OF\ AN\ NCA}.$$

The probabilities of staying in an NCA and moving from NCA to a gene are calculated in an analogous fashion. For practical purposes, let's skip the math and suppose we found the following transition probabilities:



The model above is called a Markov Chain. Notice something important: even though a Markov Chainl is able to capture the sequential nature of a genome, the next state only depends on our current state.

Let's keep going. Suppose, for the sake of illustration, that our whole genome is $G = (AUC)$. We know the likelihoods of each of these characters being in a gene. And we know the general probability of any position being a GENE or an NCA. So we can calculate the probability of a sequence of states. For instance, this whole genome could be one long GENE. Which means we start at state GENE, and stay there for two iterations. We have all the pieces to use Bayes's theorem to calculate the probability that the whole genome is a GENE. We can also calculate the probability that we start in an NCA, move to a GENE, and then go back to an NCA. If we calculate the probability of every single sequence of states given our observed genome, we can choose the most likely one.

This is what we call a **Hidden Markov Model (HMM)**. The states of the genome: $\{GENE,\ NCA\}$, are hidden from us, but by exhaustively applying Bayes's Theorem to each and every possible sequence of states, we choose the most probable one.

One last kink to fold: our example genome has only three letters, making calculating every sequence of steps pretty easy. But what happens when the genome has millions of nucleotides? Do we still exhaustively calculate the probability of every possible sequence of states? And it's good to keep in mind that real bioinformatics applications would likely consider much more states than just $\{GENE,\ NCA\}$.

Unfortunately, yes. To be sure that we're choosing the most likely sequence of events, we have to calculate the probability of all of them. There is some good news though: we can use a trick to significantly speed up our computations. Dynamic Programming. We encode the problem in matrix form and store the partial results (the optimal path to every subpath of the genome), which allows us to solve the problem in exponential time.

So after running a dynamic programming algorithm, we're ready to start selecting all the regions of the virus's genome that encode proteins. Now it's just a matter of finding the right one, which means: back to the lab; there's still a whole lot of tests to be done. But we're a lot closer to finding out a vaccine than we were before. A *lot* closer: by finding out all the genes with our probabilistic model we probably saved several years' worth of laborious lab work.

I hope you now have some appreciation for Bayes's Theorem and have at least some notion of how we use Hidden Markov Models in bioinformatics. Thanks for reading.