

BNFO301: Introduction to Bioinformatics

Calculation of probability of rare events

Perhaps the prime directive to genome analysis is *Go forth and discover meaning*. But how do you find meaning in a bunch of letters that make up a genome? As you'll recall from one of our [previous adventures](#), Fuller et al (1984) noted four instances of the sequence TTAT[CA]CA[CA]A in the 540-bp region identified as the origin of replication of *E. coli*. Suppose that you didn't know that this sequence is important or even that the 540-bp region has an interesting function. Would you be surprised to find four instances of that sequence in the 540 basepairs?

This is what life is like in studying genomes. You look for a pattern in what at first seems like noise. You find a pattern and are so happy, you want to celebrate. But before you get too ecstatic, try this:

SQ1. Display the first 100 nucleotides of the genome of *Prochlorococcus marinus* Med4 (also known as *Prochlorococcus marinus* subsp *Pastoris* CCMP1986), nicknamed *med4* in [CyanoBIKE](#) and [CCMP1986 in PhAnToMe/BioBIKE](#). Do you see anything striking about it? (Hint: Look at the T's. See any clusters?)

That's pretty remarkable, no?

I. How to calculate the probability of multiple rare occurrences (Part 1)

Or *is* it remarkable? By "remarkable", I mean something well beyond what you'd expect by chance. So, we have to know, what *would* you expect by chance?

SQ2. Modifying SQ1, what is the probability of finding TTTT at a specific position?

You might quickly crank out $(1/4)*(1/4)*(1/4)*(1/4)$, arguing that starting at that specific position, n ,

n	$n+1$	$n+2$	$n+3$
T	T	T	T

the probability of T at coordinate n is $1/4$, the probability of T at coordinate $n+1$ is also $1/4$, and so on, and you just have to multiply the four probabilities together to get the joint probability that all four positions contain T.

SQ3. What assumptions are you making in calculating the probability in this way? Are they valid?

Well, you're way beyond that level of sophistication now. Leaving aside the question of whether the four positions are *independent* of one another,¹ you wouldn't dream of presuming that all four nucleotides are equally likely.²

SQ4. Determine the appropriate nucleotide frequencies for this calculation and recalculate the probability of TTTT at a specific position.

You could readily calculate those frequencies using the tools you know from [What is a Gene](#), but an easier way to do it is using the BioBIKE function GC-FRACTION-OF. This gives you the

¹ If you don't understand the significance of "independent", then see the presentation [Probability and Genomes](#).

² Of course you wouldn't. For one thing, you've completed Problem Set 1, problem 3.

fraction of a sequence or set of sequences that are either G or C. From this, you can calculate the frequencies of all four nucleotides.

But wait a second! What sequence(s) should you give GC-FRACTION-OF? You can ask for the GC-FRACTION of the entire Med4 genome, but perhaps it's significant that the possibly remarkable clustering of TTTT in the first 100 nucleotides occurred in a region that was not a gene. Does that make a difference?

SQ5. Calculate the GC-FRACTION-OF the entire Med4 genome and also the GC-FRACTION-OF just the INTERGENIC-SEQUENCES-OF Med4 (making use of the BioBIKE function of that name). Is there any significant difference in the nucleotide frequencies of the intergenic sequences relative to those of the entire genome?

OK. Now you have possibly appropriate nucleotide frequencies and can calculate the probability of a single instance of TTTT at a specific position. Call this probability p . But p isn't good enough -- you didn't see just one instance.

SQ6. What's the probability of finding all three of the TTTT sequences you observed in the first 100 nucleotides?

You might think that the probability of one instance is p , so the probability of three simultaneous instances must be p^3 . You'd have a stronger argument for that position if you were calculating the probability of TTTT at coordinate 14 AND coordinate 23 AND coordinate 78. But – be honest – you'd have been equally surprised if you had found TTTT at, say, coordinates 29, 47, and 88 or at any other three coordinates. There are *lots* of sets of three positions that would have surprised you. Lots... Hmmm, we need a bit more precision here.

SQ7. Exactly how many ways are there of finding three instances of TTTT in 100 nucleotides? For example, you could find them at coordinate 1, coordinate 2, and coordinate 3. Or you could find them at coordinate 1, coordinate 2, and coordinate 4. And so forth. How many possible starting coordinates are there? Within that set, how many subsets of three coordinates are there?

In other words, how many ways are there of picking three coordinates from 97 of them? This, of course, is high school math – combinations. But who amongst us remembers the equation for calculating combinations? I'm talking about "*How many ways are there of picking k objects from N total objects*"?

II. Aside: How to remember combinations

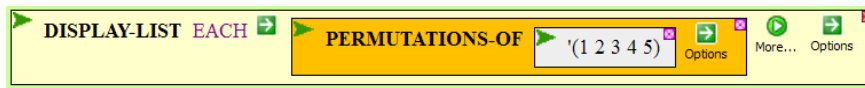
You're going to remember combinations if I can help it, since you'll certainly have ample need for them in genome analysis. Let's consider a specific instance, one you can do by hand: How many ways are there of picking 2 objects from 5 objects? Never mind the math, just count them.

SQ8. List in a systematic way all the ways of picking two numbers from the set (1 2 3 4 5), without repetitions. How many pairs did you get?

If you like, you can check yourself by going into CyanoBIKE and using the COMBINATIONS-OF function. It may be convenient to display the result, one result per line:



This is fine for cases where you can write out all of the combinations, but usually this isn't practical. You're going to have to find a way to calculate it. Here's how, using the case where you know the answer. First, calculate all the ways you can permute the set (1 2 3 4 5). To see how many there are, it might help again to write out all the possibilities. BioBIKE can help here:



SQ9. How many permutations are there of that set of five numbers?

Please don't count them. And don't look up a formula either. Instead, look at the numbers and *see* the formula. I'll help you with **Fig. 1A**, which shows only part of the listing of the permutations. If you look at the full listing, you'll see that the first column consists first of the group of permutations starting with 1, then those starting with 2, and so forth.

SQ10. How many such groups are there?

Focus on one of the groups. It doesn't matter which one. I chose the group that begins with 2. Within that group, there is a subgroup that begins with 1, another that begins with 3, and so forth.

SQ11. How many such subgroups are there?

Focus on one of the *sub*-subgroups. I chose the one beginning with 3. It has three sub-sub-subgroups, each of which has 2 sub-sub-sub-subgroups, each of which has 1 line.

SQ12. From this analysis, how many total lines are there?

5 groups, each with 4 subgroups, each with 3 sub-subgroups, each with 2 sub-sub-subgroups, each with 1 line. Simple multiplication (or more compactly, 5!) and you're there.³

You can readily move from there to the number of combinations of two elements from 5 elements, as shown in **Fig 1B**, which shows all the possible combinations in the leftmost red box, all 5! of them.

...no, wait, there aren't as many as 5!, because many are listed twice. Consider (2 1). It's listed as many times as there are permutations of (3 4 5). And (2 3) is listed as many times as there are permutations of (1 4 5).

SQ13. How many times is each pair listed? How many permutations are there of 3 elements?

If I collapse each group, for example collapsing ((2 1 3 4 5) (2 1 3 5 4)...) to a single line (2 1 (3 4 5)) and do the same with all of the groups, I get what's shown in **Fig 2**.

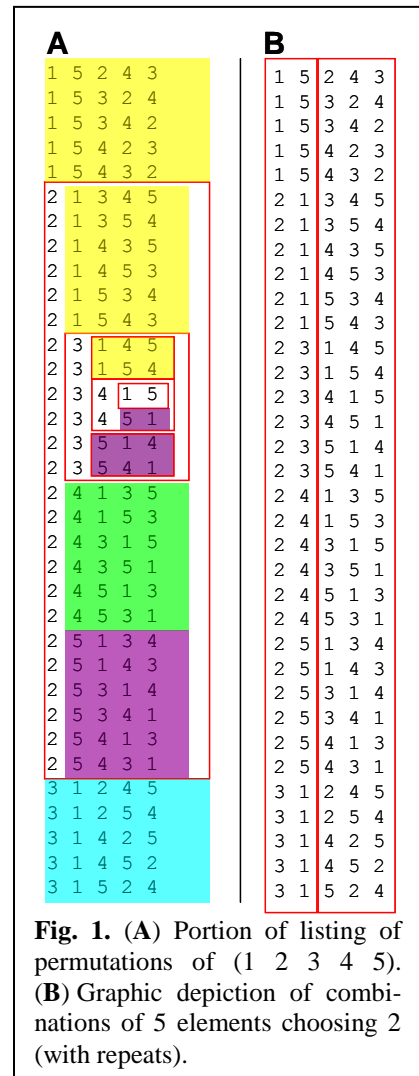


Fig. 1. (A) Portion of listing of permutations of (1 2 3 4 5). (B) Graphic depiction of combinations of 5 elements choosing 2 (with repeats).

³ This strategy predates genomic analysis by [a few thousand years](#).

SQ14. How many lines are there in Fig. 2. Use that number and the fact that there used to be 3! copies of each line before the collapse to recalculate the total number of lines in Fig. 1.

But this isn't the end either, because there is still much double counting. For example, how many times does (1 2) appear (bearing in mind that (1 2) is no different from (2 1)). In fact every pair is listed two times. Why two times? Because there are that many permutations of two elements. Cross out the duplicates and you end up the true number of combinations

In brief, if you're interested in the number of combinations of 5 elements choosing two of them at a time, then you get them by listing the total number of permutations (5!) and dividing by the permutations of the two elements (2!) and the elements that are left (3!).

SQ15. From this, come up with a general formula to obtain the combinations of N elements, choosing k of them at a time.

Sure you know the formula, but take the time to say it while thinking of the *process* we went through, so that the formula becomes a summary of that process.

1 2	(3 4 5)	3!
1 3	(2 4 5)	3!
1 4	(2 3 5)	3!
1 5	(2 3 4)	3!
2 1	(3 4 5)	3!
2 3	(1 4 5)	3!
2 4	(1 3 5)	3!
2 5	(1 3 4)	3!
3 1	(2 4 5)	3!
3 2	(1 4 5)	3!
3 4	(1 2 5)	3!
3 5	(1 2 3)	3!
4 1	(2 3 5)	3!
4 2	(1 3 5)	3!
4 3	(1 2 5)	3!
4 5	(1 2 3)	3!
5 1	(2 3 4)	3!
5 2	(1 3 4)	3!
5 3	(1 2 4)	3!
5 4	(1 2 3)	3!

Fig. 2. Combinations of 5 elements choosing 2, partially collapsed to remove repeats.

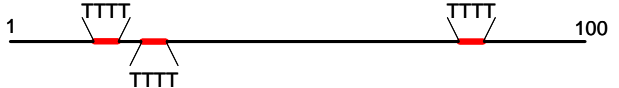
III. How to calculate the probability of multiple rare occurrences (Part 2)

Back to our problem. We calculated the probability of finding TTTT at a specific position – a small number. What we want is the probability of finding TTTT at three positions, any three positions within 100 nucleotides. At least in a naïve way, we know how to calculate the probability of finding TTTT at three *specific* positions and figure that we'd be home by multiplying this number by the number of combinations of three instances in the 100 nucleotides. Let's continue with this naïve analysis with the tools developed in **Section II**.

SQ16 (=SQ7). Exactly how many ways are there of finding three instances of TTTT in 100 nucleotides? For example, you could find them at coordinate 1, coordinate 2, and coordinate 3. Or you could find them at coordinate 1, coordinate 2, and coordinate 4. And so forth. How many possible starting coordinates are there? Within that set, how many subsets of three coordinates are there?

SQ17. Why did I claim (at the end of Section I) that there were 97 possible coordinates in 100 where you could find TTTT? Consider, if there were only 4 nucleotides, how many possible coordinates would there be?

So it seems we have our answer:

$$\text{Probability of } \begin{matrix} 3 \text{ TTTT in} \\ 100 \text{ bp} \end{matrix} = \frac{97!}{3! (97-3)!} p^3$$


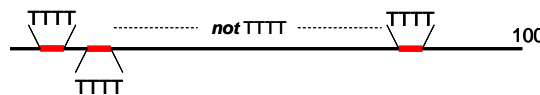
You can read this as the probability of one placement of the three red TTTT's times the number of possible placements there are. ...but wait a second! I said this is the probability of three TTTT's, but what about the black regions in between the red regions. There's nothing in my expression that excludes the possibility that another TTTT might sneak in. So it's not clear what

my expression means, but it's *not* the probability of precisely 3 TTTT's. To get that I have to add another factor that ensures that there are no additional TTTT's.

SQ18. If p is the probability that a given coordinate begins TTTT, then what is the probability that a given coordinate does *not* begin TTTT?⁴

SQ19. Suppose that 3 and only 3 positions begin TTTT. Then how many of the 97 total positions do *not* begin TTTT? Using your answer to SQ18, what is the probability that *all* of those positions simultaneously to not begin TTTT?

So now we have:

$$\text{Probability of precisely 3 TTTT in 100 bp} = \frac{97!}{3! (97-3)!} p^3 (1-p)^{94}$$


The diagram shows a horizontal line representing a 100 bp sequence. Three segments are highlighted in red, each labeled 'TTTT' with a bracket underneath. The first and third segments are at the beginning and end of the sequence, respectively. The middle segment is also labeled 'TTTT' with a bracket underneath. Dotted lines between the segments represent the remaining 94 bp, with the text 'not TTTT' written above them. The number '100' is at the far right end of the line.

SQ20. Generalize this expression, replacing 97 with N (the total number of sites), 3 with k (the total number of hits), so that the expression contains no numbers, only symbols.

Understand that there are many problems with this expression. For one thing, it matters what the sequence is. If TTTT is at coordinate 14, then another instance may be at coordinate 15, overlapping by three nucleotides. In fact, it's way more likely to be at coordinate 15 than some distant coordinate. However, if TTAA is at coordinate 14, then there is no possibility of TTAA appearing at coordinate 15. So there aren't necessarily 97 possible coordinates. Some, but not all, of the problems go away as p , the probability of getting a hit gets smaller and smaller.

The expression above is called a binomial coefficient, because it is one term in the expansion of $(x + y)^N$, where in this case $x = p$, and $y = (1-p)$.

IV. Practical calculation of the probability of multiple rare occurrences: Poisson expression

We now have a general expression that works best when p , the probability of a hit,

$$\text{Probability of precisely } k \text{ hits in } N \text{ bp} = \frac{N!}{k! (N-k)!} p^k (1-p)^{N-k}$$

where p , as usual, is the probability of one hit. As it stands, this equation is not very useful, as it doesn't work well with large values for N , and when you're dealing with genomes, N will often have *very* large values. The equation won't be very useful so long as it contains N . Fortunately, it can be simplified in a few steps that whisk N away, and only one step requires anything more than middle school math (that step requires the first semester of calculus).

First, p is replaced by a different entity that is sometimes easier to measure, the *expected number of hits*, λ , where $\lambda = N p$. This should make sense... here, let's try it.

SQ21. If each lottery ticket has a 1 in a million chance of success, and 10 million people buy lottery tickets, what is the expected number of winners? Cast this problem in terms of λ , N , and p .

Substituting for p :

⁴ If you don't understand how to do this, then see the presentation [Probability and Genomes](#).

$$\begin{aligned} \text{Probability of} \\ \text{precisely } k \text{ hits} \\ \text{in } N \text{ bp} \end{aligned} = \frac{N!}{k! (N-k)!} (\lambda/N)^k (1 - \lambda/N)^{N-k} = \frac{N!}{k! (N-k)!} \frac{\lambda^k}{N^k} (1 - \lambda/N)^{N-k}$$

Now we come to the first trick: simplifying some of those factorials:

$$\begin{aligned} N! &= N \cdot (N-1) \cdot (N-2) \cdot \dots \cdot ((N-k)+1) \cdot (N-k) \cdot ((N-k)-1) \cdot ((N-k)-2) \cdot \dots \cdot 2 \cdot 1 \\ &= N \cdot (N-1) \cdot (N-2) \cdot \dots \cdot ((N-k)+1) \cdot (N-k)! \end{aligned}$$

(and since N is way bigger than k :⁵)

$$\begin{aligned} &\overset{1}{\approx} N \cdot \overset{2}{N} \cdot \overset{3}{N} \cdot \dots \cdot \overset{k}{N} \cdot (N-k)! \\ &= N^k \cdot (N-k)! \end{aligned}$$

Replacing $N!$

$$\begin{aligned} \text{Probability of} \\ \text{precisely } k \text{ hits} \\ \text{in } N \text{ bp} \end{aligned} = \frac{N^k/(N-k)!}{k! (N-k)!} \frac{\lambda^k}{N^k} (1 - \lambda/N)^{N-k} = \frac{\lambda^k}{k!} (1 - \lambda/N)^{N-k}$$

Almost all the N 's have disappeared! Our focus turns to the last one. Simplifying $(1 - \lambda/N)^{N-k}$ may not look very promising, until you recall a famous word problem:

Suppose you have a bank account that pays you an interest of r and that interest is compounded 2 times a year. What is the effective interest? What if it is compounded 4 times a year? 12 times a year? 365 times a year? Continuously?

Such problems are solved with the following equation:

$$\begin{aligned} \$ \text{ at end of year} &= \text{principle} (1 + r / 1) \quad [\text{if not compounded}] \\ &= \text{principle} (1 + r / 2)^2 \quad [\text{if compounded twice a year}] \\ &= \text{principle} (1 + r / 4)^4 \quad [\text{if compounded four times a year}] \\ &= \text{principle} (1 + r / n)^n \quad [\text{if compounded } n \text{ times a year}] \end{aligned}$$

Using L'Hospital's rule (or a calculator), you can show that as n goes to infinity, the expression has a limit of:

$$\$ \text{ at end of year} = \text{principle} \cdot e^r \quad [\text{if compounded continuously } (n \rightarrow \text{infinity})]$$

Note that the key factor in the problem is the same form as the problematic factor in our expression, so long as r is taken to be $-\lambda$ and N and $N-k$ go to infinity. The factor can therefore be replaced by $e^{-\lambda}$! This gives the final result:

$$\begin{aligned} \text{Probability of} \\ \text{precisely } k \text{ hits} \\ \text{in } N \text{ bp} \end{aligned} = \frac{\lambda^k}{k!} e^{-\lambda}$$

To remind you, k is the number of hits (in our case 3), and λ is the expected number of hits (in our case 97 p). It's all easy for a calculator to handle because there are no huge numbers anywhere.

SQ22. Using this equation, calculate the probability of finding precisely 3 TTTT's in the first 100 nucleotides of Prochlorococcus.

⁵ Plug in some numbers and you'll see that it's true, like $N = 10^6$ and $k = 3$.

V. Who needs math? Getting answers by computer simulation

There may be days where you don't remember compound interest or how to calculate combinations. It happens to everyone. In such cases, consider turning to your friend, the computer. If you can create a plausible computer model for your problem, then let it replace the math.

SQ23. Recall your goal: You want to know what is the probability of observing what you in fact observed – 3 instances of TTTT's in the first 100 nucleotides of Prochlorococcus – but in 100 random nucleotides instead. If you're going to model this on the computer, what specific tasks will you need to be able to accomplish?

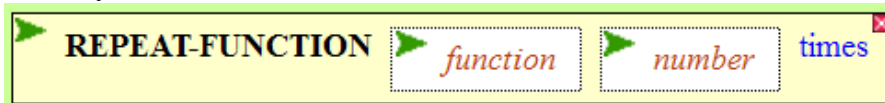
BioBIKE gives you tools that make simulating this sort of problem very easy. First, you're going to need 100 random nucleotides. Go to the ALL menu and bring into the workspace the RANDOM-DNA function, then execute it. In a second you'll get a sequence in the Result pane (though not all of it will be visible).

SQ24. How does this sequence compare to the first 100 nucleotides of Prochlorococcus. You'll want to look at least at its length and its nucleotide composition.

The notion of "random DNA sequence" is surprisingly difficult to define. The sequence you're looking at has approximately the same frequencies of nucleotides. As such, it's nothing at all like the Prochlorococcus genome, and it would be a very poor random DNA sequence to use for comparison. Perhaps you want a random sequence that matches the nucleotide frequencies of the Prochlorococcus genome. Or maybe matches those of the intergenic sequences of Prochlorococcus. Or perhaps those of this particular intergenic sequence or this particular 100-nt region. For the sake of simplicity, let's say you want a random sequence like the one you were looking at, i.e. the first 100 nucleotides of the genome. Now look at the Options for RANDOM-DNA and note with glee that there is an option called LIKE. If you give it a sequence or set of sequences (or even an entire organism), it will create a piece of DNA the same length and nucleotide composition but different sequence.

SQ25. Do an experiment, getting the COUNT-OF the number of TTTT's in a RANDOM-DNA sequence LIKE the first 100 nucleotides of Prochlorococcus. How many did you count? Try it again... now how many?

Since the sequence you're examining is made at random each time, you may get a different answer each time you run the function. The last step is to repeat the experiment many times and get a COUNT-OF how many times you get 3 instances of TTTT. The REPEAT-FUNCTION function allows you to do this:



Drag the function you made to count TTTT's in the random sequence into the *function box*. Then specify in the *number box* how many times you want to do the experiment. Then execute the function.

SQ26. How close does your experimental value come to the theoretical value you calculated in the last section?

SQ27. How remarkable do you think is the occurrence of the TTTT's?