What is a Gene?

A BioBIKE Tour

II. What is the beginning of a gene?

II.A. Overview of the problem

So it turns out to be quite simple to recognize the beginning of a gene. You just look for a change in color. But wait! Cells can't see color! And they don't understand coordinate systems! What do they do instead? Fortunately, standard textbooks have a ready answer (Fig. 2). They all seem to agree that genes begin with an initiation codon AUG (or ATG if looking at the DNA) and end with a stop codon.

"The codon AUG, which specifies Met (methionine), is also the "start" codon for polypeptide synthesis." [Hartl & Jones (1999). *Essential Genetics]*

"...the first codon to be translated -- the initiation codon -- is an AUG set in a special context at the 5' end of the gene's reading frame..." [Hartwell et al (2000). *Genetics: From Genes to Genome]*

"One of these codons, AUG… is the first codon read in an mRNA in translation…" [Russell et al (2008), *Biology, the Dynamic Science,* p.306]

Fig. 2. Where do genes start? A sampling of the lore found in textbooks.

Nonetheless, it couldn't hurt to take a look ourselves, since we happen to have a genome handy. If you put aside what you already know, how could you figure out from a genome's worth of genes how genes begin? Imagine a string of a few million nucleotides. Somewhere within it are a few thousand genes. How does a cell pick out where they begin?

No idea? Well, imagine instead a book of a few million letters (that's a very long book!). Somewhere in it are a few thousand sentences (those are very long sentences!). How do ***you*** pick out where they begin?

That's not so difficult a question to answer. If you examine your internal processes, you may come up with the following two strategies:

a. Look for an internal cue -- i.e. a capital letter. Words with capital letters are candidates to begin sentences.
b. Look for an external cue -- i.e. punctuation. Words following periods or question marks are candidates to begin sentences.

Maybe these strategies will work for identifying the beginnings of genes.

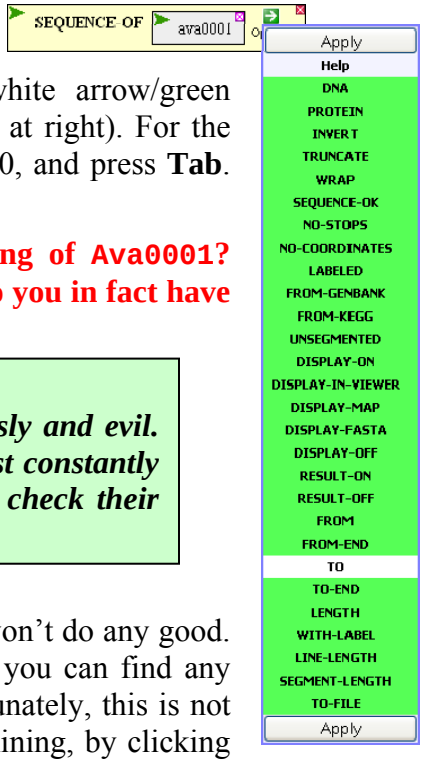
7. Do these strategies work for English sentences?

7a. Blank your mind of all knowledge of English and use the two strategies to identify the beginnings of sentences in the first paragraph of Section II. Did they work? What further rules could you devise to make them work better?

II.B. Where is the information determining the beginning of a gene? (Part I)

Let's first try the first strategy, examining the beginnings of genes to see if anything jumps out as interesting.

8. To focus on just the gene's beginning, modify the SEQUENCE-OF function again, this time adding the TO option. Do this by mousing over the Options Icon (white arrow/green background) and clicking TO and then APPLY (see figure at right). For the TO *value*, click the word *value*, type the first coordinate, 10, and press **Tab**. Then execute the function.



8a. What nucleotides do you expect from the beginning of `Ava0001`? Compare what you see with what you saw before. Do you in fact have the first 10 nucleotides of `Ava0001`?

> ***Words to live by***
> ***Computers are a source of power, but they're also sly and evil. They will fool you every time they can, and you must constantly be on your guard. Any time you have a chance to check their work by hand, DO IT.***

9. You can stare all you like at a single gene, but it probably won't do any good. What you need to do is to look at ***many*** genes and see if you can find any general features (like the equivalent of capital letters). Fortunately, this is not difficult to do. Delete the name of the gene you were examining, by clicking the delete icon (see right). Then click the entity box, but don't type anything in it. Instead, go to the GENOME button in the function palette and click GENES-OF, producing:

This says you want the sequence of ***all*** the genes up to the 10th nucleotide. All the genes of what? Certainly *Avar* can provide a healthy number of genes, so click the *entity* box, type *Avar*, and press Enter. Then execute the function.

You should have been given a new window with sequences. Scroll down to see what you have. Do you have sequences for all the genes? BioBIKE routinely saves you from being buried under huge mounds of output. If you ***really*** want to see the whole thing, you can, by changing your **preferences** through the FILE button. But you don't really want to.

9a. What fraction of the genes of *Avar* would you estimate are represented in your output?

9b. Consider the results and -- something only humans can do -- find something interesting, a pattern of undefined nature you think is significant.

10. No doubt you see something, a general tendency that holds for only the first three nucleotides. Let's focus on them. Modify the SEQUENCE-OF command and execute it, to generate a list of not the first ten but the ***first three nucleotides*** of each gene in *Avar*.

10a. Time to come up with a hypothesis. What single three-nucleotide sequence looks like it might serve as the capital letter (or perhaps *a* capital letter) that marks the beginning of sentences?

10b. Can you formulate a rule that will predict the beginning of every gene?

II.C. Testing the hypothesis (Part I)

By now you have come up with a fairly nice hypothesis, if you do say so yourself. But it's just a bunch of words unless you can test it, preferably with quantitation.

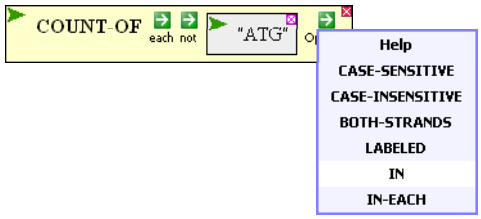
11. How could you test your hypothesis, at least with respect to *Avar*? Consider, you formed your hypothesis just looking at a small fraction of the genes. Maybe they're not representative of the whole. Maybe you're being fooled by a weird subset.

    Suppose your hypothesis has something to do with a specific sequence – let's take "ATG" as a random example, just to have something specific to talk about. You'd very much like to know what fraction of the genes of *Avar* begin with "ATG".

    11a. What two things do you have to know to calculate that fraction?

12. You can get a count of the number of genes that begin with "ATG" within the list of all gene beginnings by using the **COUNT-OF** function. Get it by mousing over the LIST-TABLES button, then LIST-ANALYSIS, and click COUNT-OF.

    The COUNT-OF box asks you to fill in a *query*? What could it want? What do you want to count? In our example, you want to count how many times "ATG" appears. Type that into the *query* box (and don't forget the quotation marks). But is that enough for a meaningful question? If I walked up to you and asked how many "ATG"s are there? Would you be able to answer the question? How many are there ***in what***?

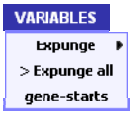
    To complete the question, mouse over the Options icon and select the IN option. Then fill it in… with what?

13. You want to get a count of "ATG"s in the list you produced in #10. There are many ways to explain this to COUNT-OF, but perhaps the cleanest is to give that list a name, by defining a variable that contains that result. To do this, mouse over the DEFINITION button in the function palette and click DEFINE. You're asked to provide two items: the name of the *variable* and the *value* of the variable. The name can be anything you like (so long as the name does not contain spaces), but it's better to be descriptive. Suppose you chose `gene-starts`. Type that in the *var* box and press **Tab**.

    The *value* is what was produced by the function you made in #10. Drag that entire function over to the *value* box. Do this by clicking on the word SEQUENCE-OF and holding the click.[‡] A ghost box will appear and a cursor. Drag the cursor to the *value* box so the cursor

[‡] If you see a red dotted outline around the SEQUENCE-OF box, that means you selected the box, probably by a brief click rather than by holding down the mouse button. Click the box again to get rid of the outline, and try again.

hovers over the box and causes it to be outlined in red. Then release the ghost box. Finally execute the DEFINE function.



If all is well, then a new VARIABLES button will appear on your palette. If you mouse over it, you should see the variable you just created.

14. Return to the still incomplete COUNT-OF function, click its gray *value* box, then go to the VARIABLES button and click your new variable to bring it into the box. Finally execute the function. A number should appear in the purple Results pane at the bottom of the screen.

    14a. What does the number mean? Use it in an English sentence.

    14b. What else do you need to know in order to find the fraction of genes in *Avar* that begin with "ATG"?

15. In #4a you estimated the number of genes in *Avar*. You could use that number, but you can do much better. What you want is a COUNT of the number of genes in the organism. Bring down a fresh version of the COUNT-OF function (as you did in #12). This time, however, fill its *query* box with GENES-OF *avar*, and execute the function.

    15a. Now, finally, you can answer the question from #11: What fraction of genes in *Avar* begin with "ATG"?[§]

    15b. What might the rest start with? Examine the window you generated in #9.

16. Modify the COUNT-OF box from #14, replacing "ATG" with other candidate start triplets.

    16a. Add up all the counts. Do they add up to the total number of genes? How many genes are not accounted for?

II.D. What are the exceptions?

17. So a ***lot*** of genes begin "ATG", and some others begin with triplets similar to "ATG", but there must be other genes that don't begin this way. Why are they so special?

    17a. Can you identify any gene of *Avar* within the list you generated in #9 that is one of the exceptions?

    17b. Estimate how long it would take to go through the entire list, presuming you figured out how to display the entire thing.

Words to live by

The greatest insights are often gained by investigating exceptions to rules. Are they truly exceptions? Do they help us understand what the rule really means? Can you think of a more powerful rule that encompasses the exceptions?

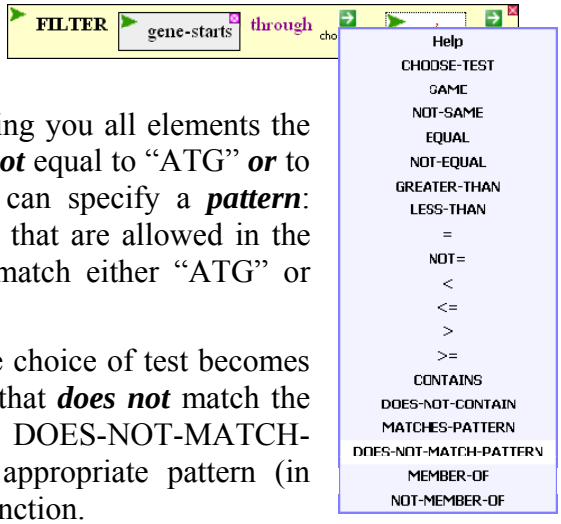
18. A long, repetitive task… sounds like something for a computer. But computers aren't genies ("Bring me fame, fortune, and happiness!"). They'll only do precisely what you tell them to. You need to imagine with as much detail as possible, what you want the computer to look for and to do when going through the list of triplets.

[§] Don't have a calculator? An estimate will do. But actually you DO have a calculator – you're staring at one! And it's WAY more powerful than the little box you might routinely use. Investigate the ARITHMETIC button in the palette.

18a. What instructions can you come up with?

> ***Words to live by***
> ***Many go through their scientific lives primarily in tool-driven mode, asking "I know how to do X, what question can I answer?" Alternatively, there's goal-driven mode, "I need to answer question X… What tool will get me there? Does it exist? If not, how can I build it?" Both modes have their uses, the first for rapid progress, the second for more profound progress.***

19. Maybe someone has already built the tool you need already. Mouse over the LISTS-TABLES button, then the LIST-EXTRACTION option, and click on FILTER. The FILTER box that comes down to your workspace asks for three things: *data* to be filtered, a *test* to be applied to each element in the list, and the *value* for the test. The data is clear enough: that would be the list of triplets you may have named *gene-starts*.

    The test and value are more problematic. If you were looking for all of the "ATG"s, you could specify the test to be EQUAL (or the more relaxed SAME) and the value to be "ATG", giving you all elements the same as "ATG". But you want to find elements ***not*** equal to "ATG" ***or*** to other triplets similar to "ATG". For that you can specify a ***pattern***: "[*abc*…]TG" where [*abc*…] specifies the letters that are allowed in the first position. For example, "[AC]TG" would match either "ATG" or "CTG".

    If you understand the notion of patterns, then the choice of test becomes clear. You want the filter to allow every triplet that ***does not*** match the pattern. Mouse over **choose-test** and choose DOES-NOT-MATCH-PATTERN. Then type in the *value* box the appropriate pattern (in quotation marks), press **Enter**, and execute the function.

    19a. Look at the result pane. How many triplets were you expecting to see? Recall your answer to 16a.

    19b. How many triplets in fact are there? Do they match your specification (i.e. *not* matching the pattern)? You can count them by hand, but why bother? Use COUNT-OF, dragging the result you just got into the COUNT-OF's *query* box.

20. These triplets don't look anything like the others you've seen. ***Why???*** What genes are they attached to? Why are those genes so special? Life would be so much easier if each triplet were labeled by the name of the gene it was attached to. Then you can look up the genes to see what they have in common.

    It turns out such labeling is not difficult. Go back to the box where you defined the list of triplets. Mousing over the **Options** icon in the orange SEQUENCE-OF box (don't be waylaid by the same icons in the other boxes), you'll see an option LABELED. Click on it, and re-execute the DEFINE box. You'll see in the Result pane that each triplet has been associated with the name of its gene. To make that easier to see, bring down a DISPLAY box from the INPUT-OUTPUT menu and drag the result from the Result pane into DISPLAY's *object* box. Then execute. Take a brief look at the output window that pops up. Don't throw it away. It will become more interesting in a moment.

Now that you've redefined *gene-starts*, re-execute FILTER, producing a list of weird triplets and the genes to which they're attached. DISPLAY this list as you did the previous one.

20a. Compare the two displays. What is special about those genes that don't begin with ATG or similar? Why the correlation?

20c. Come up with a new hypothesis, one that incorporates what you now know and explains the beginnings of genes.

20d. How does your hypothesis accord with what appears in textbooks (Fig. 2 above).

Please understand that Hartl, Hartwell, Russell, and the rest are not ignoramuses. They know perfectly well what you've just discovered. But a textbook is not the place for details and nuance. If textbooks went much beyond generality, they would grow to several times bigger than what they already are. If you want the truth, go to nature.

Please understand as well that you did ***not*** go to nature You trusted that what a computer claimed to be the beginning of a gene really ***is*** the beginning of a gene. How does ***it*** know?

20e. If you can't trust textbooks (and you can't), and you can't trust computers (and you *really* can't), what *can* you trust?

Supplemental Problems

P5. That last question is important enough to repeat. What *can* you trust?

P6. To continue, how could you determine to a greater degree of satisfaction what is the real beginning of Ava0001?

P7. Consider the first three nucleotides of the genes that *don't* begin with ATG or similar. Do they look like random triplets? What explanation do you have for the nonrandomness? You might want more information, like what are the functions of the genes containing these triplets. You can get descriptions of functions using the DESCRIPTION-OF function (GENES-PROTEINS menu). Enter the name of a gene into the *entity* box and execute.

Or, if you're in a hurry, you can get descriptions for all of the genes all at once, by filling the *entity* box with NONCODING-GENES-OF (found in the GENOME menu, GENOME-ELEMENTS submenu), and filling its entity box with *avar*. If you execute this, you will no doubt be disappointed to read in the Results pane that it is a list (well, it is a list, isn't it?). You want a description of each element of the list, so click the Each icon within NONCODING-GENES-OF, and while you're in the area, click the DISPLAY option for easier reading. It should look something like this: