(Updated 21 Jan 2011)

What is a Gene?

A BioBIKE Tour

III. What determines the beginning of a protein-encoding gene?

III.A. Overview of the problem

You now know that we must distinguish between genes that are transcribed to messenger RNA that encode proteins and those that are transcribed to RNA (like transfer RNA and ribosomal RNA) that do not encode proteins. You may be pleased with having gone beyond the textbook and discovering the pattern of triplets that begin protein-encoding genes Perhaps you are satisfied that you've achieved a glimpse into the mind of God…

…but wait a second! What if those triplets are red herrings! Certainly they are ***correlated*** with the beginnings of genes, but do they actually ***determine*** the beginnings? After all, capital letters don't always indicate the beginnings of English sentences.

21. Suppose for the moment that "ATG" or similar is sufficient to mark the beginning of a gene. That means that anywhere in the genome a cell found an "ATG", it would think, "A gene starts here."

    21a. Accept all that for a moment. How many "ATG"s should there be in the entire genome relative to the number of genes?

22. You already know how many genes are in *Avar*… How many "ATG"s does it have? You can answer this by modifying a COUNT-OF box so that "ATG"s are counted IN *Avar*.

    22a. Consider the proposition: "ATG is sufficient to mark the beginning of a gene." Support or refute that statement using evidence from your computational experiments. Be sure to present your reasoning, a prediction based on that reasoning, and the actual result you obtained.

We seem to have a problem. Section II indicated that protein-encoding genes always begin with "ATG" or some variant (at least in *Anabaena variabilis*) (at least if we accept CyanoBIKE's determinations of where genes begin). But now you have found that "ATG" is not sufficient to determine the beginning of a gene. There must be some other source of information used by a cell.

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

Evidently every "ATG" doesn't signify the start of a gene. That's not so strange. Every capital letter doesn't signify the start of a sentence. We make sense of English texts by noting capital letters in light of what precedes them. Perhaps it would be rewarding to do the same with genes.

How to do that? Old news. You examined the first three nucleotides of all the genes of *Avar* through the use of SEQUENCE-OF, specifying the range of interest. Why not do the same with the sequences ***before*** the first nucleotide? As an example, consider the sequence of *Avar* that you got a long time ago in Section I, Step 1. Look at the 15 nucleotides preceding the start of Ava0001. How can we extract those nucleotides from that gene? How can we do the same for ***all*** the genes of *Avar*?

23. BioBIKE numbers nucleotide coordinates like so (using the sequence of Ava0001 as an example):

-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	**+1**	**+2**	**+3**
C	A	A	A	G	G	T	G	G	T	A	A	A	G	G	**A**	**T**	**G**

Notice that there is no coordinate 0, just as there's no year 0 in the Gregorian calendar. You obtained the first three nucleotides by asking for the SEQUENCE-OF the gene FROM 1 TO 3 (FROM 1 is the default so was not necessary). To get the 15 nucleotides preceding the gene (called the 15 upstream nucleotides), you can use the same function, specifying FROM -15 TO -1. Use SEQUENCE-OF to get those nucleotides, and verify that you got the right ones.

24. Now that you know how to do this for one gene, get the upstream sequence from all the genes of *Avar*. You generalized SEQUENCE-OF over all genes before, you will recall. This time, however, confine your attention to those genes that encode proteins. To do this, note that by mousing over the GENOME button and then the GENOME-ELEMENTS sub-menu, you'll see a provocative function called CODING-GENES-OF. Bring it into SEQUENCE-OF and execute the function when it is complete.

24a. Can you find a generality from amongst all the sequences upstream from the coding genes of *Avar*?

25. It's possible that you ***can*** see such a pattern, but it's by no means obvious. You probably will need some help. Statistics often provide help. Suppose you went through every upstream sequence and counted how many nucleotides at the -15th position were A's, how many were C's, etc. Then you did the same thing for the -14th position, and so forth. I've tabulated Ava0001 below (leaving only about 6000 genes to go!).

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
A		1	1	1							1	1	1		
C	1														
G					1	1		1	1					1	1
T							1			1					

(compare the table in #23 to make sure I counted correctly). When completed, you will have a ***P***osition-***S***pecific ***S***coring ***M***atrix (PSSM) for the upstream sequences of *Avar*'s protein-encoding genes. I do not advise you actually do this unless you have a few weeks spare time, but for a computer, it's child's play.

How? First DEFINE a variable that contains the upstream sequences, as you did in Section II, #13. Use a different name in the *var* box of course, something appropriate like *upstream-sequences*. If you execute the definition, the variable's name should appear under your VARIABLES button

26. Second bring down from the STRINGS-SEQUENCES menu, BIOINFORMATIC-TOOLS submenu, the function called MAKE-PSSM-FROM. Into the *aligned-list* box, take the name of the variable you just made from the VARIABLES by clicking on the name. Executing the MAKE-PSSM-FROM function will produce in the Results pane a 2-dimensional table.

27. Display the table through the DISPLAY-TABLE function, found on the INPUT-OUTPUT menu. Drag the result from #26 into the *table* box and execute the function.

There are a few disconcerting features of the displayed table that differ from what you might have expected from the table shown in #25. First of all, the nucleotides are on top rather than on the left. This is easily remedied by choosing INVERT from the **Options** icon of DISPLAY-TABLE and re-executing the function.

The second unwanted feature is that the columns go from 1 to 15 rather than -15 to -1. This is because PSSM started counting from 1. No one told it that the first column was -15. You'll need to do the translation in your mind.

The third feature may be the most confusing. The table obviously does not present counts.

27a. What sense can you make of the fractions presented in the table?

If you don't see any meaning right off, try adding up the numbers of each column.

28. You can ponder the numbers, but sometimes it's easier to see relationships when they are graphically displayed. Go back to the INPUT-OUTPUT button and click the PLOT function. Then drag into the *data-list* box the same result (i.e. from #25) that you dragged a moment ago. Execute the function. You should get two new windows displayed. The upper window is the legend for the plot that appears in the lower window.

28a. Identify which of the four lines is associated with which of the four nucleotides. This isn't easy, owing to the quality of the graph. It might help to look at the graph and the table from #27 at the same time.

28b. What tendencies do you see at different positions upstream from protein-encoding genes of *Avar*?

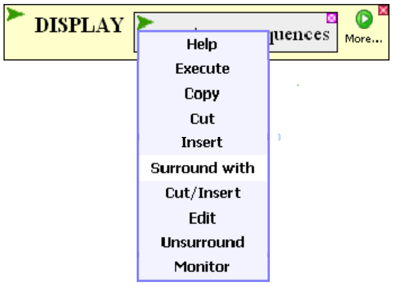
28c. Re-examine the displayed results from #24 and see if you can find hints of a pattern.

III.C. Testing the hypothesis (Part II)

It's all a trick! The human mind can find patterns in anything! How do you know what you saw is not just a product of your overactive imagination? To put it in a way that's more testable, how do you know that a pattern as deviant as the one you observed could not have arisen from the analysis of random sequences? If it could, then it is unlikely to have biological significance. It is remarkably easy to put the matter to a test.

29. First, let's get a good idea of what the upstream sequences really look like. Bring down a DISPLAY function (from the INPUT-OUTPUT menu), or better, reuse the one you already have on your screen. To do this, just click the red x icon in the upper right corner of DISPLAY's argument box to get rid of what is already there, then click the box and fill it with your variable, *upstream-sequences*, found on the VARIABLES button. Executing the function lists all the upstream sequences.

30. Now randomize those sequences by surrounding *upstream-sequences* with the RANDOMIZE function. To do this, mouse over the Action Icon of *upstream-sequences* and click SURROUND-WITH, as shown to the right. A red dotted line will surround the object box.



31. Mouse over the STRINGS-SEQUENCES menu, then the STRING-PRODUCTION submenu, and click on RANDOMIZE, causing the RANDOMIZE function to surround *upstream-sequences*. If you executed the function now, the list of sequences would be randomized, i.e. the order of the list's elements. That's not what you want. You want to randomize each sequence within the list. Click the EACH-IN option of RANDOMIZE. Now Execute the DISPLAY function.

    31a. Carefully examine the first sequence in the display of *upstream-sequences* and compare it to the first sequence in the display of the randomized *upstream-sequences*. What is the relationship of one sequence to the other? Check this by comparing the second and third sequences.

    31b. Execute the DISPLAY function a second time and compare the new display to the old. What is the relationship of one to the other?

32. Now that you know what RANDOMIZE does, return your attention to the MAKE-PSSM-FROM function and do the same trick, surrounding *upstream-sequences* with RANDOMIZE (making sure to click the EACH-IN option). Execute the new MAKE-PSSM-FROM function and drag the result into PLOT. Execute the new PLOT function.

    32a. What tendencies do you see at different positions upstream from protein-encoding genes of *Avar* when the upstream sequences are randomized?

    32b. Execute PLOT again, eliciting a new randomization. What tendencies do you see now?

Words to live by

The world (aided by your imagination) is constantly trying to convince you of things that aren't true. Statistics provide us with a defense, but it is a rare person who really understands how to use statistical tools with understanding. They are often misused. Consider, however, that statistical tests generally answer the question "Could X have arisen by chance?" You can often answer this question in a more intuitive way by making chance work for you and seeing what results.

By repeatedly plotting the PSSMs generated from the randomized sequences, you can get a feel for what kind variation may be obtained by chance.

32c. Argue for or against the proposition that the pattern you observed in #28 is attributable to chance. Whatever argument you make, USE NUMBERS!