What is a Gene?

A BioBIKE Tour

I. Preliminaries

I.A. Framing the question

What is a gene? If you're hoping to find an answer here like "the unit of inheritance" or something like that, move on! A few years ago, 25 researchers were locked in a room to find a definition and came up with "*A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions.*" Wha-a-at? That's what happens when humans argue about human definitions.

This tour is not about how humans define a gene but what ***cells*** actually use to determine where to find the information necessary to make a protein. There are millions of nucleotides (letters) in a bacterial genome, billions in the human genome. How does a cell sift through this to find the beginning and the end of the information?

You might imagine genes within a cell to be something like what is depicted in Fig. 1A. In this model, the beginning and end of a gene is just the beginning and end of the squiggle – no problem. Alternatively, you might imagine what is depicted in Fig. 1B, genes embedded within a single piece of DNA. Which model is closer to the truth?

How to find out? Find a text book? Go to an authority figure? Google? Why trust any of them when you can go so much closer to the source – ***Ask the genome!***

That's what this tour is about: finding things out for yourself.

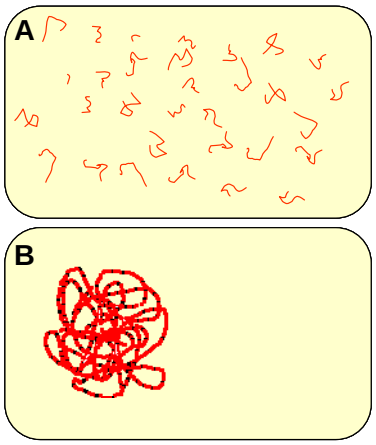


Fig. 1: Physical nature of genes. Which is it? (**A**) Genes as independent pieces of DNA floating within a cell? Or (**B**) Genes as elements of a chromosome?

I.B. Accessing genomes through BioBIKE

Ask the genome? How? Allow me to introduce BioBIKE, a web-based compendium of knowledge about genes, genomes, and metabolism and a graphical programming interface. You can get there by using Firefox (no other browser is supported) to go to:

> http://biobike.csbc.vcu.edu/ (the BioBIKE portal)
> Click the link to the CyanoBIKE VCU mirror
> Log in as yourself (no spaces or special characters)
> Click New Session (you will at some point have to allow popups from this site)
> (If this is the first time you've logged in) Answer a few questions and click Register

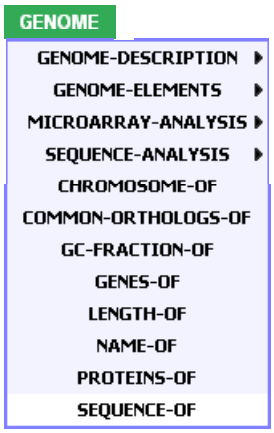
Note that you are greeted by an invitation to a tour of the BioBIKE interface. It will tell you about the conventions used by the interface – what the various icons mean, how to cut and paste, etc. If you want to skip it for now, fine. If you change your mind, you can always find it by mousing over the red HELP button and clicking Interface Tour. The most critical lessons are these:

- Functions (the things you want to do) are found through the buttons at the top of the screen – the palette.
- A function doesn't do anything until you ***execute*** it, by double clicking the name of the function in its box.
- Enter things by clicking on gray entry boxes (turning the box white). Entry is not deemed complete until you press **Enter** or **Tab** (turning the box gray). A function cannot be executed if any inner box is white.

I.C. Understanding the genome coordinate system

Back to the question – are genes independent entities or are they embedded in larger entities? Enough abstractions… let's look at specific genes in the specific genome of a specific organism. I've chosen the cyanobacterium *Anabaena variabilis*… what? You've never heard of it? You're missing an important story! They're amongst the closest living relatives to the progenitor of chloroplasts, directly or indirectly responsible for the support of most life on earth. They're also amongst the only creatures able to live in pure salt water, using sunlight for energy, the air for carbon and nitrogen, and water for electrons. If you like (or if you value your typing fingers) you can call it by its nickname, *Avar*.

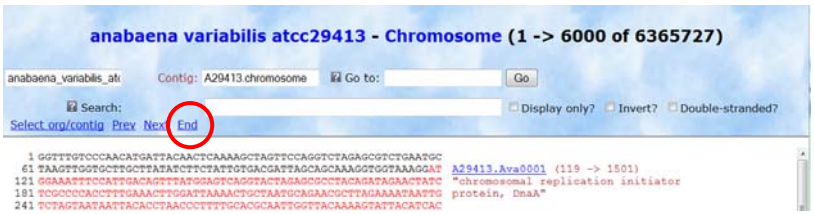
1. What does its genome look like? To examine the DNA sequence of *Avar*, mouse over the **GENOME** button on the function palette and click **SEQUENCE-OF**, as shown to the right. This will bring into the workspace the following box:



   Click on the word *entity* and type `Avar` in the white box,* and press **Enter**. Now that the function is complete, execute it, either by double-clicking the word **SEQUENCE-OF** or by mousing over the action icon (the green wedge) of the function and clicking **Execute**.

   If all has gone well, a new window should appear, a ***sequence viewer*** showing the genome of *Avar*. This DNA sequence is supposed to contain all the information required to define everything the organism is and can do. Amazing, isn't it?

2. If you scroll through the sequence, you'll find that it seems to stop after only a few thousand nucleotides. "Only"? Well, yes. Is just a few thousand letters enough to define something as complicated as an organism? Is that all there is? If you look at the top of the window, you'll see something like:



Click the **End** link, bringing you to what is evidently the end of the chromosome.

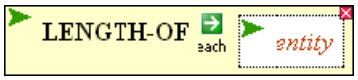
* If you forget the name of the organism, you can get it from a list, by mousing over the **DATA** button, then **organisms**, then the name of the organism. But typing is easier. By the way, upper/lower case doesn't matter.

2a. What is the coordinate of the last nucleotide in the genome?

2b. Go back to the beginning by clicking Start. How many nucleotides are shown? Approximately what fraction is that of the whole?

BioBIKE protected you from having the full 6.4 million nucleotide sequence dropped on your screen. That's the nature of genomes. They're too big for a mere human to take in. We'll have to sometimes look at small pieces of it and sometimes consider it abstractly as a whole.

3. Check that 6.4 million figure. From the green CyanoBIKE screen, GENOME button, bring down the LENGTH-OF function. Click the *entity* box and type *Avar*, then press **Enter** and execute the function (remember the double-click trick). The result of the function will appear in the purple Result pane at the bottom of the page.



3a. Does the result from LENGTH-OF agree with your answer to Problem 2a? Any ideas? If you ask a question twice and get two different answers, that's generally a sign that you're really asking two different questions. Try to find clues that indicate how the two questions you asked differ from each other.

4. Go back to the sequence window and notice that portions of the DNA sequence are given in different colors and that to the right there are names and descriptions of these regions – genes!.

4a. Judging by the number of genes in the portion of the genome you can see (after scrolling), how many genes do you think are in the entire genome? Do an appropriate calculation. Hold on to this number, we'll check it later.

5. Notice also that there are numbers next to the name of the gene. For example, there appears "(119 → 1501)" next to Ava0001.

5a. Consider those numbers. What could they mean? Test your hypothesis using the sequence and the numbers at the left of each line. (Yes, this means counting, but not much)

6. Check your hypothesis in another way. Return to the CyanoBIKE screen and edit the SEQUENCE-OF function, replacing *Avar* with the gene *AVA0001*.[†] You can edit the function by clicking on the entry box, changing its color from gray to white. Press **Enter** to close the entry box, and then execute the function.

6a. Compare the sequence that appears in the window with the sequence that you obtained in step 1. What is the relationship?

6b. Of the two models shown in Figure 1 above, which is supported by CyanoBIKE?

[†] It doesn't matter whether you type the name of a gene or organism with upper or lower case, so do whatever makes you happiest.

Supplemental Problems

P1. Use the coordinates for Ava0001 as indicated in the displayed sequence of *Avar*, calculate the length of the gene. Check your answer by using the LENGTH-OF function, providing *Ava0001* as the entity.

P2. Extract the sequence of Ava0001 from the *Avar* chromosome, using SEQUENCE-OF, providing *Avar* as the entity, and using the FROM and TO options of SEQUENCE-OF (mouse over the Options Icon or, if you can't find them, sneak a peak at Section 2 #8 of this tour). Compare the sequence you obtain to the sequence of the gene you obtained in #6. Are they the same?

P3. In the displayed sequence of *Avar*, go to the last nucleotide of the sequence. You did this before by clicking End. This time do it by typing the coordinate of the last nucleotide in GoTo box, then click Go. The sequence you get is quite different from what you got by clicking End. How do you explain the result?

P4. What are the first three nucleotides of Ava0003? Determine this in at least two ways: (1) from the display of the *Avar* chromosome, and (2) from the display of the sequence of Ava0003. Do your answers agree? If they don't, why not? It may be useful to note that the sequence viewer has multiple display formats.